

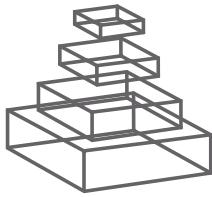
frontiers RESEARCH TOPICS

INTRINSIC MOTIVATIONS AND
OPEN-ENDED DEVELOPMENT IN
ANIMALS, HUMANS, AND ROBOTS

Topic Editors

Gianluca Baldassarre, Tom Stafford,
Marco Mirolli, Peter Redgrave,
Richard Michael Ryan and Andrew Barto





FRONTIERS COPYRIGHT STATEMENT

© Copyright 2007-2015
Frontiers Media SA.
All rights reserved.

All content included on this site, such as text, graphics, logos, button icons, images, video/audio clips, downloads, data compilations and software, is the property of or is licensed to Frontiers Media SA ("Frontiers") or its licensees and/or subcontractors. The copyright in the text of individual articles is the property of their respective authors, subject to a license granted to Frontiers.

The compilation of articles constituting this e-book, wherever published, as well as the compilation of all other content on this site, is the exclusive property of Frontiers. For the conditions for downloading and copying of e-books from Frontiers' website, please see the Terms for Website Use. If purchasing Frontiers e-books from other websites or sources, the conditions of the website concerned apply.

Images and graphics not forming part of user-contributed materials may not be downloaded or copied without permission.

Individual articles may be downloaded and reproduced in accordance with the principles of the CC-BY licence subject to any copyright or other notices. They may not be re-sold as an e-book.

As author or other contributor you grant a CC-BY licence to others to reproduce your articles, including any graphics and third-party materials supplied by you, in accordance with the Conditions for Website Use and subject to any copyright notices which you include in connection with your articles and materials.

All copyright, and all rights therein, are protected by national and international copyright laws.

The above represents a summary only. For the full conditions see the Conditions for Authors and the Conditions for Website Use.

Cover image provided by Iblb sarl,
Lausanne CH

ISSN 1664-8714

ISBN 978-2-88919-372-1

DOI 10.3389/978-2-88919-372-1

ABOUT FRONTIERS

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

FRONTIERS JOURNAL SERIES

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing.

All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

DEDICATION TO QUALITY

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.

Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view.

By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

WHAT ARE FRONTIERS RESEARCH TOPICS?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area!

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: researchtopics@frontiersin.org

INTRINSIC MOTIVATIONS AND OPEN-ENDED DEVELOPMENT IN ANIMALS, HUMANS, AND ROBOTS

Topic Editors:

Gianluca Baldassarre, Italian National Research Council, Italy

Tom Stafford, University of Sheffield, United Kingdom

Marco Mirolli, Istituto di Scienze e Tecnologie della Cognizione, Italy

Peter Redgrave, University of Sheffield, United Kingdom

Richard Michael Ryan, University of Rochester, USA

Andrew Barto, University of Massachusetts Amherst, USA

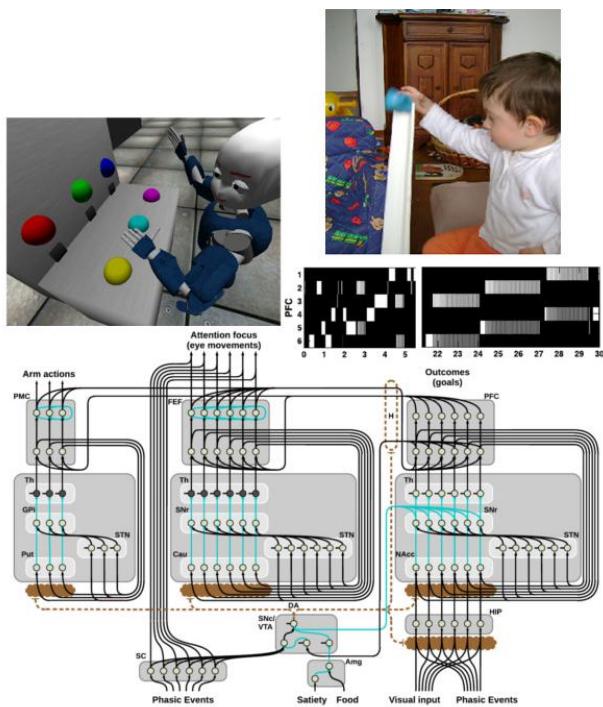


Image by Gianluca Baldassarre, based on: Fiore, V. G.; Sperati, V.; Mannella, F.; Mirolli, M.; Gurney, K.; Firston, K.; Dolan, R. J. & Baldassarre, G. (2014). Keep focussing: striatal dopamine multiple functions resolved in a single mechanism tested in a simulated humanoid robot. *Frontiers in Psychology*, 5 (124).

The aim of this Research Topic for *Frontiers in Psychology* under the section of Cognitive Science and *Frontiers in Neurorobotics* is to present state-of-the-art research, whether theoretical, empirical, or computational investigations, on open-ended development driven by intrinsic motivations. The topic will address questions such as: How do motivations drive learning? How are complex skills built up from a foundation of simpler competencies? What are the neural and computational bases for intrinsically motivated learning? What is the contribution of intrinsic motivations to wider cognition?

Autonomous development and lifelong open-ended learning are hallmarks of intelligence. Higher mammals, and especially humans, engage in activities that do not appear to directly serve the goals of survival, reproduction, or material advantage. Rather, a large part of their activity is intrinsically motivated - behavior driven by curiosity, play, interest in novel stimuli and surprising events, autonomous goal-setting, and the pleasure of acquiring new competencies. This allows the cumulative acquisition of knowledge and skills that can later be used to accomplish fitness-enhancing goals. Intrinsic motivations continue during adulthood, and in humans artistic creativity, scientific discovery, and subjective well-being owe much to them.

The study of intrinsically motivated behavior has a long history in psychological and ethological research, which is now being reinvigorated by perspectives from neuroscience, artificial intelligence and computer science. For example, recent neuroscientific research is discovering how neuromodulators like dopamine and noradrenaline relate not only to extrinsic rewards but also to novel and surprising events, how brain areas such as the superior colliculus and the hippocampus are involved in the perception and processing of events, novel stimuli, and novel associations of stimuli, and how violations of predictions and expectations influence learning and motivation.

Computational approaches are characterizing the space of possible reinforcement learning algorithms and their augmentation by intrinsic reinforcements of different kinds. Research in robotics and machine learning is yielding systems with increasing autonomy and capacity for self-improvement: artificial systems with motivations that are similar to those of real organisms and support prolonged autonomous learning. Computational research on intrinsic motivation is being complemented by, and closely interacting with, research that aims to build hierarchical architectures capable of acquiring, storing, and exploiting the knowledge and skills acquired through intrinsically motivated learning.

Now is an important moment in the study of intrinsically motivated open-ended development, requiring contributions and integration across a large number of fields within the cognitive sciences. This Research Topic aims to contribute to this effort by welcoming papers carried out with ethological, psychological, neuroscientific and computational approaches, as well as research that cuts across disciplines and approaches.

Table of Contents

- 06 *Intrinsic Motivations and Open-Ended Development in Animals, Humans, and Robots: An Overview***
Gianluca Baldassarre, Tom Stafford, Marco Mirolli, Peter Redgrave, Richard M. Ryan and Andrew Barto
- 11 *Novelty or Surprise?***
Andrew Barto, Marco Mirolli and Gianluca Baldassarre
- 26 *Learning Autonomy in Two or Three Steps: Linking Open-Ended Development, Authority, and Agency to Motivation***
Tjeerd C. Andringa, Kirsten A. van den Bosch and Carla Vlaskamp
- 44 *Emergent Structured Transition From Variation to Repetition in a Biologically-Plausible Model of Learning in Basal Ganglia***
Ashvin Shah and Kevin N. Gurney
- 60 *Modeling Effects of Intrinsic and Extrinsic Rewards on the Competition Between Striatal Learning Systems***
Joschka Boedecker, Thomas Lampe and Martin Riedmiller
- 72 *Keep Focussing: Striatal Dopamine Multiple Functions Resolved in a Single Mechanism Tested in a Simulated Humanoid Robot***
Vincenzo G. Fiore, Valerio Sperati, Francesco Mannella, Marco Mirolli, Kevin Gurney, Karl Friston, Raymond J. Dolan and Gianluca Baldassarre
- 89 *No Learning where to go without First Knowing where You're Coming From: Action Discovery is Trajectory, not Endpoint Based***
Martin Thirkettle, Tom Walton, Peter Redgrave, Kevin Gurney and Tom Stafford
- 98 *Robust Active Binocular Vision through Intrinsically Motivated Learning***
Luca Lonini, Sébastien Forestier, Céline Teulière, Yu Zhao, Bertram E. Shi and Jochen Triesch
- 108 *The Role of Intrinsic Motivations in Attention Allocation and Shifting***
Dario Di Nocera, Alberto Finzi, Silvia Rossi and Mariacarla Staffa
- 123 *Novelty, Attention, and Challenges for Developmental Psychology***
Emily Mather
- 127 *Autonomous Visual Exploration Creates Developmental Change in Familiarity and Novelty Seeking Behaviors***
Sammy Perone and John P. Spencer
- 148 *Image Free-Viewing as Intrinsically-Motivated Exploration: Estimating the Learnability of Center-of-Gaze Image Samples in Infants and Adults***
Matthew Schlesinger and Dima Amso

- 160 Which is the Best Intrinsic Motivation Signal for Learning Multiple Skills?**
Vieri G. Santucci, Gianluca Baldassarre and Marco Mirolli
- 174 PowerPlay: Training an Increasingly General Problem Solver by Continually Searching for the Simplest Still Unsolvable Problem**
Jürgen Schmidhuber
- 188 Confidence-Based Progress-Driven Self-Generated Goals for Skill Acquisition in Developmental Robots**
Hung Ngo, Matthew Luciw, Alexander Förster and Jürgen Schmidhuber
- 207 Linear Combination of One-Step Predictive Information with an External Reward in an Episodic Policy Gradient Setting: A Critical Analysis**
Keyan Zahedi, Georg Martius and Nihat Ay
- 218 Incremental Learning of Skill Collections Based on Intrinsic Motivation**
Jan H. Metzen and Frank Kirchner
- 230 Neural Model for Learning-To-Learn of Novel Task Sets in the Motor Domain**
Alex Pitti, Raphael Braud, Sylvain Mahé, Mathias Quoy and Philippe Gaussier
- 245 Curiosity Driven Reinforcement Learning for Motion Planning on Humanoids**
Mikhail Frank, Jürgen Leitner, Marijn Stollenga, Alexander Förster and Jürgen Schmidhuber
- 260 A Psychology Based Approach for Longitudinal Development in Cognitive Robotics**
J. Law, P. Shaw, K. Earland, M. Sheldon and M. H Lee
- 279 A Game Theoretic Framework for Incentive-Based Models of Intrinsic Motivation in Artificial Systems**
Kathryn E. Merrick and Kamran Shafi
- 296 Imitation Learning Based on an Intrinsic Motivation Mechanism for Efficient Coding**
Jochen Triesch
- 304 Self-Organization of Early Vocal Development in Infants and Machines: The Role of Intrinsic Motivation**
Clément Moulin-Frier, Sao M. Nguyen and Pierre-Yves Oudeyer
- 324 A Motivation Model for Interaction Between Parent and Child Based on the Need for Relatedness**
Masaki Ogino, Akihiko Nishikawa and Minoru Asada
- 335 From Self-Assessment to Frustration, a Small Step Toward Autonomy in Robotic Navigation**
Adrien Jauffret, Nicolas Cuperlier, Philippe Tarroux and Philippe Gaussier



Intrinsic motivations and open-ended development in animals, humans, and robots: an overview

Gianluca Baldassarre^{1*}, Tom Stafford², Marco Mirolli¹, Peter Redgrave², Richard M. Ryan³ and Andrew Barto⁴

¹ Laboratory of Computational Embodied Neuroscience, Institute of Cognitive Sciences and Technologies, National Research Council, Rome, Italy

² Department of Psychology, University of Sheffield, Sheffield, UK

³ Department of Clinical and Social Sciences in Psychology, University of Rochester, River, New York, USA

⁴ Department of Computer Science, University of Massachusetts Amherst, Massachusetts, USA

*Correspondence: gianluca.baldassarre@istc.cnr.it

Edited and reviewed by:

Eddy J. Davelaar, Birkbeck College, UK

Keywords: intrinsic motivations, novelty and surprise, cumulative learning and development, computational models, autonomous robotics, reinforcement learning, brain and behavior, review

1. INTRODUCTION

This editorial article introduces the Frontiers Research Topic and Electronic Book (eBook) on *Intrinsic Motivations* (IMs), which involved the publication of 24 articles with the journals *Frontiers in Psychology – Cognitive Science* and *Frontiers in Neurorobotics*. The main objective of this Frontiers Research Topic is to present state-of-the-art research on IMs and open-ended development from an interdisciplinary perspective involving human and animal psychology, neuroscience, and computational perspectives. We first introduce in this section the main themes and concepts on IMs from different interdisciplinary perspectives. These themes and concepts have been reviewed more extensively in other works (e.g., see Barto et al., 2004; Oudeyer and Kaplan, 2007; Mirolli and Baldassarre, 2013; Barto, 2013), but they are briefly reported here both to meet the needs of the reader new to the field and to introduce the concepts and terms we use in the succeeding sections. In the next four sections, we give an overview of the Topic contributions grouped by four themes. A final section draws the conclusions.

Autonomous development and lifelong open-ended learning are hallmarks of intelligence. Higher mammals, and especially humans, engage in activities that do not appear to directly serve the goals of survival, reproduction, or material advantage. Rather, many activities seem to be carried out “for their own sake” (Berlyne, 1966), play being a prime example, but including other activities driven by curiosity and interest in novel stimuli or surprising events. Autonomous setting goals and working to acquire new forms of competence are also examples of activities that often do not confer obvious evolutionary benefit. Activities like these are thus said to be driven by intrinsic motivations (Baldassarre and Mirolli, 2013a). IMs facilitate the cumulative and virtually open-ended acquisition of knowledge and skills that can later be used to accomplish fitness-enhancing goals (Singh et al., 2010; Baldassarre, 2011). IMs continue during adulthood, and they underlie several important human phenomena such as artistic creativity, scientific discovery, and subjective well-being (Ryan and Deci, 2000b; Schmidhuber, 2010).

IMs were proposed within the animal literature to explain aspects of behavior that could not be explained by the dominant theory of motivation postulating that animals work to reduce physiological imbalances (Hull, 1943). The term “intrinsic motivation” was first used to describe a “manipulation drive” hypothesized to explain why rhesus monkeys would engage with mechanical puzzles for long periods of time without receiving extrinsic rewards (Harlow et al., 1950). Other studies showed how animal instrumental actions can be conditioned with the delivery of apparently neutral stimuli: for example, monkeys were trained to perform actions to gain access to a window from which they could observe conspecifics (Butler, 1953), and mice were trained to perform actions that resulted in clicks or in moving the cage platform (Kish, 1955). The psychological literature on IMs initially linked them to the perceptual properties of stimuli, such as their complexity, novel appearance, or surprising features (Berlyne, 1950, 1966). Later, IMs were also related to action, in particular to the competence (“effectance”) that an agent can acquire to willfully make changes in its environment (White, 1959). This relation of IMs with action and their effects was later linked to the possibility of autonomously setting one’s own goals (Ryan and Deci, 2000a).

Computational approaches, in particular machine learning and autonomous robotics, are concerned with IMs and open-ended development as these are thought to have the potential to lead to the construction of truly intelligent artificial systems, in particular systems that are capable of improving their own skills and knowledge *autonomously and indefinitely*. The relation of these studies with those on IMs in psychology were first highlighted by Barto et al. (2004) and Singh et al. (2005). The investigation of IMs from a computational perspective can lead to theoretical clarifications, in particular with respect to the computational mechanisms and functions that might underlie IMs (Mirolli and Baldassarre, 2013). IM mechanisms have been classified as being either *knowledge-based* or *competence-based* (Oudeyer and Kaplan, 2007): the former based on measures related to the acquisition of information, and the latter on measures related to the learning of skills. More recently,

knowledge-based IMs have been further divided into *novelty-based IMs* and *prediction-based IMs* (Baldassarre and Mirolli, 2013b; Barto et al., 2013). Novelty-based IMs are elicited by the experience of stimuli that are not in the agent's memory (e.g., novel objects, or novel object-object or object-context combinations); prediction-based IMs are related to events that surprise the agent by violating its explicit predictions.

These distinctions have been formalized in the computational models proposed in the literature. Seminal works in machine learning (Schmidhuber, 1991), later developed to function in robots (Oudeyer et al., 2007), have proposed algorithms rewarding actions that allow the agent to improve the quality of a "predictor" component with which it anticipates the effects that such actions produce on the environment. Other researchers have proposed robots capable of detecting and focussing on novel stimuli (e.g., Marsland et al., 2005), or systems capable of detecting anomalies in datasets (Nehmzow et al., 2013). Additional research threads have focussed on action and control, in particular on IMs guiding the autonomous acquisition of motor skills (Barto et al., 2004), on the decision about which of several skills to practice at any time (Schembri et al., 2007; Santucci et al., 2013), and on the the autonomous formation of goals guiding skill acquisition (Baranes and Oudeyer, 2013). Other computational mechanisms related to the idea of IMs are being proposed in the growing field of *active learning*, in particular in relation to supervised learning systems (Settles, 2010).

Recent neuroscientific investigations are revealing brain mechanisms that possibly underlie the IM systems investigated in the behavioral and computational literature. However, unfortunately such investigations are carried out under agendas different from the one on IMs, e.g., in relation to dopamine, memory, motor learning, goal-directed behavior, and conflict monitoring, so comprehensive views are still missing. A large body of research shows how the hippocampus, a brain compound system playing pivotal functions for memory, has the capacity to detect the novelty of various aspects of experience, from the novelty of single items to the novelty of item-item and item-context associations (Ranganath and Rainer, 2003; Kumaran and Maguire, 2007). This detection is then capable of triggering the release of neuromodulators, such as dopamine, that modulate the functioning and learning processes of the hippocampus itself and other brain areas, e.g., of the frontal cortex involved in higher cognition, action planning, and action execution (Lisman and Grace, 2005). Other studies have shown that unexpected stimuli can activate the superior colliculus, a midbrain structure that plays a key role in oculomotor control, which in turn causes phasic bursts of dopamine affecting trial-and-error learning processes happening in basal ganglia, a brain region known to be involved in learning to select actions and other cortex contents (Redgrave and Gurney, 2006). Dopamine signals have also been shown to have an interesting direct relationship with information seeking (Bromberg-Martin and Hikosaka, 2009). Noradrenaline, another neuromodulator targeting a large part of brain, has been shown to be involved in signaling violations of the agent's expectations (Sara, 2009). The failure (Carter et al., 1998) or success (Ribas-Fernandes et al., 2011) in accomplishing goals and sub-goals, possibly themselves set by IMs, has been shown to have neural

correlates that might affect succeeding motivation, engagement, and learning. Bio-inspired/bio-constrained computational modeling is linking some of these neuroscientific results to specific computational mechanisms, e.g., in relation to dopamine (e.g., see the pioneering work of Kakade and Dayan, 2002, and Mirolli et al., 2013) and goal-directed behavior (Baldassarre et al., 2013).

The 24 interdisciplinary contributions to the present Research Topic can be clustered into four groups. The first group of six contributions (*IMs and brain and behavior*) focuses on different types of IM mechanisms implemented in the brain. The second group of five contributions (*IMs and attention*) focuses on the role of IMs in attention. The third group of eight contributions (*IMs and motor skills*) focuses on IMs as drives for the acquisition of manipulation and navigation skills, often with an emphasis on their function in enabling cumulative, open-ended development. Finally, the fourth group of five contributions (*IMs and social interaction*) focuses on the relationship between IMs and social phenomena, a novel area of investigation of IMs that is increasingly attracting the attention of researchers.

2. INTRINSIC MOTIVATIONS, BRAIN AND BEHAVIOR

The theoretical contribution of Barto et al. (2013) argues for the importance of distinguishing between *novelty* and *surprise* on the basis of a comprehensive analysis of the computational literature related to the two. It then shows the utility of the distinction for improved understanding of brain and behavior phenomena where the two are often confused. Andringa et al. (2013) present a broad view of possible relationships between IMs and control, exploration, and agency, linking these processes to the specialization of the left and right hemispheres of the brain and showing how the interplay between these can lead to a progressive sophistication of cognition. Shah and Gurney (2014) propose a computational model that investigates how basal ganglia, modulated by IMs, can lead to a dynamical shift from noise-based exploration to repetition that can support the acquisition of both simple and more complex motor skills (in the present case, simulated reaching skills). Boedecker et al. (2013) propose a computational model based on the distinction between dorsal and ventro-medial basal ganglia regions (supporting respectively habitual and goal-directed behavior). Through the model, the authors analyze the relation between these brain regions and IMs concerning reasoning costs and the value of information. This analysis is used to account for some empirical phenomena concerning the relationship between extrinsic and IMs. Fiore et al. (2014) propose a biologically-constrained computational model that also focuses on different portions of basal ganglia. The model shows how these regions can be differentially regulated by a unique tonic dopaminergic signal, linked to both intrinsic and extrinsic motivations, on the basis of their different sensitivity to dopamine. The model, also tested with the simulated humanoid robot iCub, shows how these modulatory mechanisms can play important adaptive functions for the control of overt attention, manipulation, and goal-directed processes. Thirkettle et al. (2013) introduce the novel "Joystick experimental paradigm" developed to study intrinsically and extrinsically driven acquisition of actions. The authors demonstrate the function and effectiveness of this paradigm by presenting behavioral experiments grounded

in the neuroscientific literature and concerning the acquisition of non-trivial motor actions.

3. INTRINSIC MOTIVATIONS AND ATTENTION

The computational work of Lonini et al. (2013) builds on a previous binocular system in which an IM learning signal is generated on the basis of the capacity of the system to reconstruct images encoded with sparse-coding features. This signal guides the acquisition of attention and vergence skills by reinforcement learning. The contribution here focuses on demonstrating the robustness of the system, in particular for recovering from disturbances and for self-recalibration. Di Nocera et al. (2014) present a behavior-based architecture that uses curiosity drives to improve the attentional capabilities of a reinforcement learning robot engaged in solving simulated survival “extrinsic” tasks. Overall, the work shows the utility of IMs to improve attention and, based on this, action selection. Mather (2013) briefly reviews research related to the familiarity-to-novelty attention shift observed in babies, and, on this basis, highlights the challenges that this phenomenon poses to theories on IMs. Perone and Spencer (2013) also deal with the familiarity-to-novelty shift. In particular, the authors propose a dynamical-field model that offers an explanation of the phenomenon as emerging from the autonomous accumulation of visual experience under the guidance of novelty-based IMs. Schlesinger and Amso (2013), referring to the results of tests of both human and computational agents engaged in solving a visual-exploration task, propose that free viewing of natural images in human infants can be understood as the effect of intrinsically motivated visual exploration driven by the goal of producing predictable gaze sequences. The authors highlight the implications of their approach for understanding visual development in infants.

4. INTRINSIC MOTIVATIONS AND OPEN-ENDED DEVELOPMENT OF MOTOR SKILLS

Santucci et al. (2013) focus on the problem of which IM signals are best suited to decide which skills to learn by reinforcement learning given a set of tasks. By comparing the results of systems receiving different IM signals, they show that the best IM signals are those based on mechanisms that measure the improvement of the skill competence rather than the errors, or error improvements, of predictors of the action effects on the environment. In a theoretical machine learning contribution, Schmidhuber (2013) proposes a system that automatically invents computational problems in order to train an increasingly-general problem solver. IM signals driving learning are generated when the system finds more efficient skills to solve all the problems generated thus far. In a similar vein, Ngo et al. (2013) propose an architecture for controlling a Katana simulated and real robot interacting with a blocks-world. The system is capable of self-generating goals based on its confidence in its predictions about how the environment will react to its actions. Zahedi et al. (2013) propose the use of task-independent IMs to support task-dependent learning on the basis of the mutual information of the past and future elements of sensor streams (predictive information). The authors conclude that a combination of predictive information with external rewards is recommended only for hard

tasks to speed-up learning but at the cost of an asymptotic performance lost. Metzen and Kirchner (2013) propose a reinforcement learning model that self-generates tasks on the basis of graphs of states and selects the skills to learn on the basis of both novelty-based and prediction-based IMs. The system is tested with navigating and octopus-like simulated robots acting in continuous domains. Inspired by infant cognition, Pitti et al. (2013) present a reinforcement-learning bio-inspired gain-fields system for learning task-sets (areas of the sensorimotor space having a common underlying cause-effect structure). The system, tested in a cognitive task and with a Kinova robot arm, is capable of recognizing a given task-set as familiar and can create a new representation for it on the basis of its uncertainty and related prediction errors. Frank et al. (2014) propose a system for controlling the humanoid robot iCub that explores the state-action space on the basis of information gain maximization so as to improve the learning of the world model used for real-time motion planning. Law et al. (2014) present a schema-based memory system inspired by child early sensorimotor development for controlling the iCub robot. The system undergoes a staged learning process to acquire eye-arm reaching skills and basic manipulation skills under the guidance of novelty- and prediction-based IMs, and the progressive release of constraints focussing attention and learning on relevant experiences.

5. INTRINSIC MOTIVATIONS AND SOCIAL PHENOMENA

In a contribution based on game theory, Merrick and Shafi (2013) propose the concept of “optimally motivating incentive” for game players, and show how different instances of such an incentive (i.e., strong power, affiliation, and achievement motivation) can be used in both modeling human behavior and designing effective artificial agents. The theoretical contribution of Triesch (2013) starts from the idea of IMs serving the function of learning “efficient coding” of sensory data and proposes that imitation can emerge as the consequence of a general intrinsic drive to compress information that leads to matching one’s own actions with those of the imitated tutor. Moulin-Frier et al. (2013) propose a model of the initial staged development of speech in infants. IMs initially drive the system to learn the control of phonation, then to produce unarticulated sounds, and finally to produce proto-syllables. The model is tested with a simulator of the vocal tract, the auditory system, the agent’s motor control, and social interactions with peers. The contribution of Ogino et al. (2013) proposes a reinforcement learning model of parent-child engagement where learning signals, similar to phasic dopamine signals, are caused by both extrinsic and intrinsic information, in particular related to the presence and novelty of emotional facial expressions. Finally, Jauffret et al. (2013) propose a bio-inspired neural architecture that uses a prediction-based algorithm applied to sensorimotor contingencies to solve complex navigation tasks and is capable of asking for help in dead-lock situations.

6. CONCLUDING REMARKS

The papers of the present Research Topic testify to the existence of ample interest on the Topic issues. At the same time, they show that the literature on IMs is still characterized by a

heterogeneity of perspectives on their possible roles in cognition and behavior and on the possible mechanisms supporting them. On the one side, this heterogeneity is expected given the recency of the attempts to systematize the psychological, neuroscientific, and computational views on IMs within broad interdisciplinary frameworks. On the other side, the heterogeneity is also an indication of the richness of intrinsically motivated phenomena, of their importance for animals' cognition and behavior, and of their utility for the design of autonomous robots and intelligent machines. The richness of this topic is expected to result in a further strengthening of the research in the field over the near future.

ACKNOWLEDGMENTS

This research has received funds from the European Commission 7th Framework Programme (FP7/2007-2013), "Challenge 2—Cognitive Systems, Interaction, Robotics," Grant Agreement No. ICT-IP-231722, Project "IM-CLeVeR—Intrinsically Motivated Cumulative Learning Versatile Robots." This Frontiers Topic was accomplished as a deliverable of the IM-CLeVeR Project.

REFERENCES

- Andringa, T. C., van den Bosch, K. A., and Vlaskamp, C. (2013). Learning autonomy in two or three steps: linking open-ended development, authority, and agency to motivation. *Front. Psychol.* 4:766. doi: 10.3389/fpsyg.2013.00766
- Baldassarre, G., Mannella, F., Fiore, V. G., Redgrave, P., Gurney, K., and Mirolli, M. (2013). Intrinsically motivated action-outcome learning and goal-based action recall: a system-level bio-constrained computational model. *Neural Netw.* 41, 168–187. doi: 10.1016/j.neunet.2012.09.015
- Baldassarre, G. (2011). "What are intrinsic motivations? a biological perspective," in *Proceedings of the International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob-2011)*, eds A. Cangelosi, J. Triesch, I. Fasel, K. Rohlffing, F. Nori, P.-Y. Oudeyer, et al. (New York, NY:IEEE), E1–E8.
- Baldassarre, G., and Mirolli, M. (Eds.). (2013a). *Intrinsically Motivated Learning in Natural and Artificial Systems*. Berlin: Springer. doi: 10.1007/978-3-642-32375-1
- Baldassarre, G., and Mirolli, M. (2013b). "Intrinsically motivated learning systems: an overview," in *Intrinsically Motivated Learning in Natural and Artificial Systems*, eds G. Baldassarre and M. Mirolli (Berlin: Springer-Verlag), 1–14. doi: 10.1007/978-3-642-32375-1_1
- Baranes, A., and Oudeyer, P. (2013). Active learning of inverse models with intrinsically motivated goal exploration in robots. *Robot. Auton. Syst.* 61, 49–73. doi: 10.1016/j.robot.2012.05.008
- Barto, A. (2013). "Intrinsic motivation and reinforcement learning," in *Intrinsically Motivated Learning in Natural and Artificial Systems*, eds G. Baldassarre and M. Mirolli (Berlin:Springer-Verlag), 17–47. doi: 10.1007/978-3-642-32375-1_2
- Barto, A., Mirolli, M., and Baldassarre, G. (2013). Novelty or surprise? *Front. Psychol.* 4:907. doi: 10.3389/fpsyg.2013.00907
- Barto, A. G., Singh, S., and Chentanez, N. (2004). "Intrinsically motivated learning of hierarchical collections of skills," in *International Conference on Developmental Learning (ICDL2004)*, eds J. Triesch and T. Jebara (New York, NY:IEEE), 112–119.
- Berlyne, D. E. (1950). Novelty and curiosity as determinants of exploratory behaviour. *Br. J. Psychol. Gen. Sec.* 41, 68–80. doi: 10.1111/j.2044-8295.1950.tb00262.x
- Berlyne, D. E. (1966). Curiosity and exploration. *Science* 143, 25–33. doi: 10.1126/science.153.3731.25
- Boedecker, J., Lampe, T., and Riedmiller, M. (2013). Modeling effects of intrinsic and extrinsic rewards on the competition between striatal learning systems. *Front. Psychol.* 4:739. doi: 10.3389/fpsyg.2013.00739
- Bromberg-Martin, E. S., and Hikosaka, O. (2009). Midbrain dopamine neurons signal preference for advance information about upcoming rewards. *Neuron* 63, 119–126. doi: 10.1016/j.neuron.2009.06.009
- Butler, R. (1953). Discrimination learning by rhesus monkeys to visual-exploration motivation. *J. Comp. Physiol. Psychol.* 46:95. doi: 10.1037/h0061616
- Carter, C. S., Braver, T. S., Barch, D. M., Botvinick, M. M., Noll, D., and Cohen, J. D. (1998). Anterior cingulate cortex, error detection, and the online monitoring of performance. *Science* 280, 747–749. doi: 10.1126/science.280.5364.747
- Di Nocera, D., Finzi, A., Rossi, S., and Staffa, M. (2014). The role of intrinsic motivations in attention allocation and shifting. *Front. Psychol.* 5:273. doi: 10.3389/fpsyg.2014.00273
- Fiore, V. G., Sperati, V., Mannella, F., Mirolli, M., Gurney, K., Friston, K., et al. (2014). Keep focussing: striatal dopamine multiple functions resolved in a single mechanism tested in a simulated humanoid robot. *Front. Psychol.* 5:124. doi: 10.3389/fpsyg.2014.00124
- Frank, M., Leitner, J., Stollenga, M., Forster, A., and Schmidhuber, J. (2014). Curiosity driven reinforcement learning for motion planning on humanoids. *Front. Neurorobot.* 7:25. doi: 10.3389/fnbot.2013.00025
- Harlow, H., Harlow, M., and Meyer, D. (1950). Learning motivated by a manipulation drive. *J. Exp. Psychol.* 40:228. doi: 10.1037/h0056906
- Hull, C. L. (1943). *Principles of Behavior*. New York, NY: Appleton-century-crofts.
- Jauffret, A., Cuperlier, N., Tarroux, P., and Gaussier, P. (2013). From self-assessment to frustration, a small step toward autonomy in robotic navigation. *Front. Neurorobot.* 7:16. doi: 10.3389/fnbot.2013.00016
- Kakade, S., and Dayan, P. (2002). Dopamine: generalization and bonuses. *Neural Netw.* 15, 549–559. doi: 10.1016/S0893-6080(02)00048-5
- Kish, G. (1955). Learning when the onset of illumination is used as the reinforcing stimulus. *J. Comp. Physiol. Psychol.* 48, 261–264. doi: 10.1037/h0040782
- Kumaran, D., and Maguire, E. A. (2007). Which computational mechanisms operate in the hippocampus during novelty detection? *Hippocampus* 17, 735–748. doi: 10.1002/hipo.20326
- Law, J., Shaw, P., Kevin, E., Sheldon, M., and Lee, M. (2014). A psychology based approach for longitudinal development in cognitive robotics. *Front. Neurorobot.* 8:1. doi: 10.3389/fnbot.2014.00001
- Lisman, J. E., and Grace, A. A. (2005). The hippocampal-VTA loop: controlling the entry of information into long-term memory. *Neuron* 46, 703–713. doi: 10.1016/j.neuron.2005.05.002
- Lonini, L., Forestier, S., Teuliere, C., Zhao, Y., Shi, B. E., and Triesch, J. (2013). Robust active binocular vision through intrinsically motivated learning. *Front. Neurorobot.* 7:20. doi: 10.3389/fnbot.2013.00020
- Marsland, S., Nehmzow, U., and Shapiro, J. (2005). On-line novelty detection for autonomous mobile robots. *Robot. Auton. Syst.* 51, 191–206. doi: 10.1016/j.robot.2004.10.006
- Mather, E. (2013). Novelty, attention, and challenges for developmental psychology. *Front. Psychol.* 4:491. doi: 10.3389/fpsyg.2013.00491
- Merrick, K. E., and Shafii, K. (2013). A game theoretic framework for incentive-based models of intrinsic motivation in artificial systems. *Front. Psychol.* 4:791. doi: 10.3389/fpsyg.2013.00791
- Metzen, J. H., and Kirchner, F. (2013). Incremental learning of skill collections based on intrinsic motivation. *Front. Neurorobot.* 7:11. doi: 10.3389/fnbot.2013.00011
- Mirolli, M., and Baldassarre, G. (2013). "Functions and mechanisms of intrinsic motivations: the knowledge versus competence distinction," in *Intrinsically Motivated Learning in Natural and Artificial Systems*, eds G. Baldassarre and M. Mirolli (Berlin:Springer-Verlag), 49–72.
- Mirolli, M., Baldassarre, G., and Santucci, V. G. (2013). Phasic dopamine as a prediction error of intrinsic and extrinsic reinforcement driving both action acquisition and reward maximization: a simulated robotic study. *Neural Netw.* 39, 40–51. doi: 10.1016/j.neunet.2012.12.012
- Moulin-Frier, C., Nguyen, S. M., and Oudeyer, P.-Y. (2013). Self-organization of early vocal development in infants and machines: the role of intrinsic motivation. *Front. Psychol.* 4:1006. doi: 10.3389/fpsyg.2013.01006
- Nehmzow, U., Gatsoulis, Y., Kerr, E., Condell, J., Siddique, N., and McGinnity, M. T. (2013). "Novelty detection as an intrinsic motivation for cumulative learning robots," in *Intrinsically Motivated Learning in Natural and Artificial Systems*, eds G. Baldassarre and M. Mirolli (Berlin: Springer-Verlag), 185–207. doi: 10.1007/978-3-642-32375-1_8
- Ngo, H., Luciw, M., Forster, A., and Schmidhuber, J. (2013). Confidence-based progress-driven self-generated goals for skill acquisition in developmental robots. *Front. Psychol.* 4:833. doi: 10.3389/fpsyg.2013.00833
- Ogino, M., Nishikawa, A., and Asada, M. (2013). A motivation model for interaction between parent and child based on the need for relatedness. *Front. Psychol.* 4:618. doi: 10.3389/fpsyg.2013.00618

- Oudeyer, P., and Kaplan, F. (2007). What is intrinsic motivation? A typology of computational approaches. *Front. Neurorobot.* 1:6. doi: 10.3389/neuro.12.006.2007
- Oudeyer, P., Kaplan, F., and Hafner, V. (2007). Intrinsic motivation systems for autonomous mental development. *IEEE Trans. Evol. Comput.* 11, 265–286. doi: 10.1109/TEVC.2006.890271
- Perone, S., and Spencer, J. P. (2013). Autonomous visual exploration creates developmental change in familiarity and novelty seeking behaviors. *Front. Psychol.* 4:648. doi: 10.3389/fpsyg.2013.00648
- Pitti, A., Braud, R., Mahe, S., Quoy, M., and Gaussier, P. (2013). Neural model for learning-to-learn of novel task sets in the motor domain. *Front. Psychol.* 4:771. doi: 10.3389/fpsyg.2013.00771
- Ranganath, C., and Rainer, G. (2003). Neural mechanisms for detecting and remembering novel events. *Nat. Rev. Neurosci.* 4, 193–202. doi: 10.1038/nrn1052
- Redgrave, P., and Gurney, K. (2006). The short-latency dopamine signal: a role in discovering novel actions? *Nat. Rev. Neurosci.* 7, 967–975. doi: 10.1038/nrn2022
- Ribas-Fernandes, J. J. F., Solway, A., Diuk, C., McGuire, J. T., Barto, A. G., Niv, Y., et al. (2011). A neural signature of hierarchical reinforcement learning. *Neuron* 71, 370–379. doi: 10.1016/j.neuron.2011.05.042
- Ryan, R., and Deci, E. (2000a). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemp. Educ. Psychol.* 25, 54–67. doi: 10.1006/ceps.1999.1020
- Ryan, R. M., and Deci, E. L. (2000b). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *Am. Psychol.* 55, 68–78. doi: 10.1037/0003-066X.55.1.68
- Santucci, V. G., Baldassarre, G., and Mirolli, M. (2013). Which is the best intrinsic motivation signal for learning multiple skills? *Front. Neurorobot.* 7:22. doi: 10.3389/fnbot.2013.00022
- Sara, S. J. (2009). The locus coeruleus and noradrenergic modulation of cognition. *Nat. Rev. Neurosci.* 10, 211–223. doi: 10.1038/nrn2573
- Schembri, M., Mirolli, M., and Baldassarre, G. (2007). “Evolving internal reinforcers for an intrinsically motivated reinforcement-learning robot,” in *Proceedings of the 6th International Conference on Development and Learning*, eds Y. Demiris, D. Mareschal, B. Scassellati, and J. Weng (New York, NY: IEEE), E1–E6.
- Schlesinger, M., and Amso, D. (2013). Image free-viewing as intrinsically-motivated exploration: estimating the learnability of center-of-gaze image samples in infants and adults. *Front. Psychol.* 4:802. doi: 10.3389/fpsyg.2013.00802
- Schmidhuber, J. (1991). “A possibility for implementing curiosity and boredom in model-building neural controllers,” in *Proceedings of the International Conference on Simulation of Adaptive Behavior: From Animals to Animats*, eds J. A. Meyer and S. W. Wilson (Cambridge, MA: MIT Press/Bradford Books), 222–227.
- Schmidhuber, J. (2010). Formal theory of creativity, fun, and intrinsic motivation (1990–2010). *IEEE Trans. Auton. Mental Dev.* 2, 230–247. doi: 10.1109/TAMD.2010.2056368
- Schmidhuber, J. (2013). Powerplay: training an increasingly general problem solver by continually searching for the simplest still unsolvable problem. *Front. Psychol.* 4:313. doi: 10.3389/fpsyg.2013.00313
- Settles, B. (2010). *Active Learning Literature Survey*. Computer Sciences Technical Report 1648, University of WisconsinMadison.
- Shah, A., and Gurney, K. N. (2014). Emergent structured transition from variation to repetition in a biologically-plausible model of learning in basal ganglia. *Front. Psychol.* 5:91. doi: 10.3389/fpsyg.2014.00091
- Singh, S., Barto, A., and Chentanez, N. (2005). “Intrinsically motivated reinforcement learning,” in *Advances in Neural Information Processing Systems 17: Proceedings of the 2004 Conference*, eds L. K. Saul, Y. Weiss, and L. Bottou (Cambridge, MA: The MIT Press).
- Singh, S., Lewis, R., Barto, A., and Sorg, J. (2010). Intrinsically motivated reinforcement learning: An evolutionary perspective. *IEEE Trans. Auton. Mental Dev.* 2, 70–82. doi: 10.1109/TAMD.2010.2051031
- Thirkettle, M., Walton, T., Redgrave, P., Gurney, K., and Stafford, T. (2013). No learning where to go without first knowing where you’re coming from: action discovery is trajectory, not endpoint based. *Front. Psychol.* 4:638. doi: 10.3389/fpsyg.2013.00638
- Triesch, J. (2013). Imitation learning based on an intrinsic motivation mechanism for efficient coding. *Front. Psychol.* 4:800. doi: 10.3389/fpsyg.2013.00800
- White, R. W. (1959). Motivation reconsidered: The concept of competence. *Psychol. Rev.* 66, 297–333. doi: 10.1037/h0040934
- Zahedi, K., Martius, G., and Ay, N. (2013). Linear combination of one-step predictive information with an external reward in an episodic policy gradient setting: a critical analysis. *Front. Psychol.* 4:801. doi: 10.3389/fpsyg.2013.00801

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 14 July 2014; accepted: 19 August 2014; published online: 09 September 2014.

Citation: Baldassarre G, Stafford T, Mirolli M, Redgrave P, Ryan RM and Barto A (2014) Intrinsic motivations and open-ended development in animals, humans, and robots: an overview. *Front. Psychol.* 5:985. doi: 10.3389/fpsyg.2014.00985

This article was submitted to Cognitive Science, a section of the journal *Frontiers in Psychology*.

Copyright © 2014 Baldassarre, Stafford, Mirolli, Redgrave, Ryan and Barto. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Novelty or Surprise?

Andrew Barto^{1*}, Marco Mirolli² and Gianluca Baldassarre²

¹ School of Computer Science, University of Massachusetts Amherst, Amherst, MA, USA

² Istituto di Scienze e Tecnologie della Cognizione, Consiglio Nazionale delle Ricerche, Rome, Italy

Edited by:

Tom Stafford, University of Sheffield, UK

Reviewed by:

Karl Friston, University College London, UK

Nathan F. Lepora, The University of Sheffield, UK

***Correspondence:**

Andrew Barto, School of Computer Science, University of Massachusetts Amherst, 272 Computer Science Building, Amherst, MA 01003, USA
e-mail: barto@cs.umass.edu

Novelty and surprise play significant roles in animal behavior and in attempts to understand the neural mechanisms underlying it. They also play important roles in technology, where detecting observations that are novel or surprising is central to many applications, such as medical diagnosis, text processing, surveillance, and security. Theories of motivation, particularly of intrinsic motivation, place novelty and surprise among the primary factors that arouse interest, motivate exploratory or avoidance behavior, and drive learning. In many of these studies, novelty and surprise are not distinguished from one another: the words are used more-or-less interchangeably. However, while undeniably closely related, novelty and surprise are very different. The purpose of this article is first to highlight the differences between novelty and surprise and to discuss how they are related by presenting an extensive review of mathematical and computational proposals related to them, and then to explore the implications of this for understanding behavioral and neuroscience data. We argue that opportunities for improved understanding of behavior and its neural basis are likely being missed by failing to distinguish between novelty and surprise.

Keywords: novelty, surprise, intrinsic motivation, novelty detection, expectation

1. INTRODUCTION

Novelty and surprise play significant roles in animal behavior and in attempts to understand the neural mechanisms underlying it. They are intimately connected to sensory processing, attention, learning, and decision making. Theories of motivation, particularly of intrinsic motivation (Deci and Ryan, 1985; Baldassarre and Mirolli, 2013), place novelty and surprise among the primary factors that arouse interest and motivate exploratory or avoidance behavior. Novelty and surprise also play important roles in technology, where detecting observations that are novel or surprising is central to many applications, such as medical diagnosis, text processing, surveillance, and security. In many—perhaps most—of these studies, novelty and surprise are not distinguished from one another: the words are used more-or-less interchangeably.

However, while undeniably closely related, novelty is in fact very different from surprise. The ordinary dictionary definition of novelty refers to the quality of not being previously experienced or encountered, while surprise refers to the result of encountering something suddenly or unexpectedly. In the most abstract setting (and ignoring many subtleties with which we attempt to deal below), detecting novelty requires examining (by one means or another) the contents of memory to determine if the stimulus has or has not previously been experienced and attended to. Surprise, on the other hand, is the result of a discrepancy between an expectation and an observed actuality. This comparison of an experience with an expectation does not require examination of the contents of memory despite the fact that an expectation is clearly built on previous experience. Something can be unanticipated without being un-experienced.

To pick just two illustrations of how natural it is to blur the distinction between novelty and surprise, consider the following

quotations. Marsland (2003) writes: “Novelty detection, recognizing that an input differs in some respect from previous inputs, can be a useful ability for learning systems, both natural and artificial. For animals, the unexpected perception could be a potential predator or a possible victim.” When discussing what happens when a naked man enters a classroom, Ranganath and Rainer (2003) write: “Suffice to say, the entrance of the naked guy was a novel event in that it was unexpected and out of context.” Although this blurring is completely understandable given how closely related novelty and surprise can be and the difficulty of formalizing the concepts, we argue that the failure to clearly distinguish between novelty and surprise precludes opportunities for improved understanding of behavior and its neural basis.

The purpose of this article is foremost to remind readers of differences between novelty and surprise, to discuss how these concepts are related, and to explore the implications of this for understanding behavioral and neuroscience data. A review of all that has been written about novelty and surprise is significantly beyond the scope of this paper. Here we present an extensive review of mathematical and computational proposals related to surprise and novelty, and we discuss these proposals in terms of our common sense notions. We also point out key factors that distinguish surprise from novelty, and we argue that some of the definitions in common use are misleading, as are some of the labels applied to results of experiments by psychologists and neuroscientists.

A caveat with respect to the interpretation of empirical data is needed. The distinction between novelty and surprise critically depends on the mechanisms in play when the nervous system produces the experimental results in question. As a consequence, it is

to be expected that one cannot say with certainty whether experimental results provide evidence for novelty or for surprise when the actual mechanisms implemented by the brain are incompletely known. However, we suggest that by distinguishing novelty from surprise some existing results might be reinterpreted in a way that improves our understanding of behavior and the neural machinery that underlies it. And, even more importantly, keeping the distinction in mind may be a useful heuristic for studying the brain. Although the names used to describe results are not important, the distinction may encourage neuroscientists to ask questions such as: Is there a predictor at play? If so, where is it? What kind of predictions does it produce? On the basis of what information? Or, if there is no prediction, what are the memories that are searched for? Where are those representations stored? These are important questions that may not arise as clearly if one fails to distinguish between novelty and surprise.

This article begins with accounts of representative examples of how the words have been interpreted, first addressing surprise (Section 2) and then novelty (Section 3). For the most part, the examples in each of these sections were chosen because they provide formalizations related to each concept, although not all of them are intended to model surprise or novelty in animals. The examples are placed in either the surprise or novelty section on the basis of which word their adherents chose to associate with them. Section 4 summarizes the main features of surprise and novelty, viewing each in an idealized form that largely ignores the more complicated issues about how they are related. Section 5 takes on some of these issues by examining the relationship between less idealized views of surprise and novelty. Some of the categories in which formalisms were placed in Sections 2 and 3 are reconsidered here. Section 6 considers how an improved understanding of differences between surprise and novelty may have beneficial consequences in neuroscience, where it can serve to sharpen the interpretation of experimental results and raise useful questions for continuing research. The article ends with a brief summary and concluding remarks.

2. SURPRISE

Of the two concepts novelty and surprise, surprise is probably the easiest to characterize. There is wide agreement that surprise is an emotion arising from a mismatch between an expectation and what is actually observed or experienced (e.g., Ekman and Davidson, 1994). Since our concern here is not with the emotion of surprise but rather with the conditions that elicit it, by surprise we mean these eliciting conditions. Surprise requires a mechanism for comparing an expectation with actuality.

But what is an expectation and how is one aroused? An expectation is usually thought of as a mental representation of a stimulus or event that is aroused by some cue or set of cues that has regularly preceded that stimulus or event in the past. Alternatively, an expectation might be aroused by an inferential process that predicts the occurrence of a stimulus or event (Berlyne, 1960). According to the most straightforward view, expectations are representations of the values that some perceptual features are likely to assume in the future. However,

expectations are naturally expressed in probabilistic terms as well, where a probability distribution over the range of possible observations can be considered to be a “belief state,” a kind of expectation that can generate surprise. If an estimated probability of an observation is available to the perceiving agent when the observation is made, then the certainty of the observation can be compared to its probability of occurrence, yielding a measure of surprise. Importantly, expectations as probabilistic beliefs are usually conditioned, in the sense of being conditional on a particular state or context. This notion of expectation (which is not the same as the expectation, or expected value, in probability theory) underlies Bayesian views of surprise that we discuss in Section 2.2 below.

The psychologist D. E. Berlyne, who wrote extensively about novelty, surprise, and curiosity, used the term *incongruity* for the situation of a stimulus creating an expectation that is unfulfilled by other stimuli that occur at the same time (Berlyne, 1960). The “two-headed lady” of his example is incongruous because her extra head violates the expectations generated by the rest of her image. Berlyne regards this as a special case of surprise that does not involve the passage of time, while acknowledging that it might actually involve time because parts of the incongruous stimulus may be scanned in succession.

Surprise plays a key role in theories of learning and finds natural expression in the framework of Bayesian statistics. Here we first discuss how prominent models of associative learning represent expectations and surprise, followed by a description of a modern Bayesian theory of surprise in which expectations appear as probability distributions over classes of environment models. Then we briefly discuss closely related information-theoretic notions of surprise. We discuss these examples in some detail because they are concrete examples of how surprise has been expressed in formal terms.

2.1. SURPRISE IN ASSOCIATIVE LEARNING THEORY

Surprise plays a key role in theories of classical, or Pavlovian, conditioning. In classical conditioning experiments, conditioned stimuli (CSs) are followed after a short time by biologically significant events (such as a shock, food, etc.), called unconditioned stimuli (USs) that reflexively produce unconditioned responses (URs). Great care is taken to prevent the animal’s response to the CS from influencing the occurrence of the US (unlike instrumental conditioning experiments where a reward is contingent on the animal’s behavior). After repeated trials consisting of the CS-US sequence, the animal comes to produce a conditioned response (CR) that resembles the UR but occurs as a response to a CS. For example, an air puff to the eye (the US) elicits a reflexive eye blink (the UR). When regularly preceded by another stimulus (the CS), say a tone or a light, occurrence of the CS comes to elicit an eye blink that anticipates the US. The process is often regarded as one of learning about predictive relationships among stimuli.

What is now called Kamin blocking is the failure of an animal to learn to elicit a CR when a CS is presented to an animal as part of a compound that includes another CS that had been previously conditioned to elicit a CR (Moore and Schmajuk, 2008). Kamin thought that this might be due to the fact that the US is

no longer surprising since it is already predicted by the previously conditioned CS:

... perhaps, for an increment in an associative connection to occur, it is necessary that the US instigates some mental work on the part of the animal. This mental work will occur only if the US is unpredicted, if it in some sense surprises the animal. Thus, in the early trials of a normal conditioning experiment, the US is an unpredicted, surprising event of motivational significance and the CS-US association is formed. (Kamin, 1969, p. 293)

This idea that an organism learns only when events violate its expectations, that is, when the organism is surprised, was elaborated by Rescorla and Wagner in the most widely-known and influential model of classical conditioning (Rescorla and Wagner, 1972):

The central notion here can also be phrased in somewhat more cognitive terms. One version might read: organisms only learn when events violate their expectations. Certain expectations are built up about the events following a stimulus complex; expectations initiated by that complex and its component stimuli are then only modified when consequent events disagree with the composite expectation. (Rescorla and Wagner, 1972, p. 75)

In the associationist tradition, the Rescorla-Wagner model adjusts associative strengths of stimuli that specify how strongly each stimulus predicts the US. Each constellation of stimuli that occurs (CS) generates a *composite expectation* for the US. This composite expectation is the weighted sum of the saliences of the stimuli in the constellation, each weighted by its corresponding associative strength for the US. The model adjusts the associative strengths that specify how strongly each component cs_i of the CS present on a trial predicts the US:

$$\Delta V_{cs_i} = \alpha_{cs_i} \beta (\lambda - V), \quad (1)$$

where V_{cs_i} is the associative strength of component i of the CS and ΔV_{cs_i} is its change, α_{cs_i} is the salience of component i of the CS, β is the learning rate parameter associated with the US, λ is the asymptote for learning for the US, and V is the composite expectation for the CS. The model adjusts the associative strengths of the stimuli present on each trial up or down depending on $\lambda - V$, the difference between the composite expectation, V , and the associative strength supported by that particular US, λ , which we call the “target associative strength.”

For the sake of brevity we skip further details and the important role this model has played in the history of animal learning theory (see Schmajuk, 2008, for a review; see also Lepora et al., 2010, and Mannella et al., 2010, for two models that capture the basic brain mechanisms with which classical conditioning is implemented in, respectively, cerebellum and amygdala). The key point is that the difference, or discrepancy, $\lambda - V$, is considered to be a measure of *surprise*: a constellation of stimuli generates an expectation that is compared with what actually happens.

The Rescorla-Wagner model is an example of an error-correcting learning rule such as the Widrow-Hoff Least Mean Square learning rule (Widrow and Hoff, 1960)

and the well-known error backpropagation algorithm (Rumelhart et al., 1986), where the US corresponds to the “teaching input” or “desired output,” and $\lambda - V$ is the error guiding learning (although the error is sometimes called a teaching signal in biological models of classical conditioning, e.g., Lepora et al., 2010). Error correction is also central to the widely-used Kalman filter and related algorithms, where the error is called the “innovation” or “measurement residual” (Welch and Bishop, 1995).

Connecting the Rescorla-Wagner model to probabilistic notions of surprise is the observation that in the case where the US is represented by a binary variable with values 0 or 1, the model computes the conditional probability of the US given possible patterns of CSs (Dayan and Long, 1998). In addition, the process of error correction is related to Bayesian learning as we discuss in Section 2.2 below.

Error correction is also the basis of Temporal Difference (TD) learning (Sutton, 1988), where the error incorporates information about the *long-term* expectation of reward and not just the immediate reward. TD learning is the basis of a model of classical conditioning that elaborates the Rescorla-Wagner model (Sutton and Barto, 1990) as well as the reward-prediction-error hypothesis about the phasic activity of dopamine producing neurons in the brain (Barto, 1995; Houk et al., 1995; Schultz et al., 1997; Schultz, 1998). TD learning is not restricted to predicting reward; the role of reward can be replaced by other stimulus features, and it can be generalized to networks of interrelated predictions (Sutton and Tanner, 2004).

In accord with the associationist view, the associative strengths of the stimuli needed for determining a composite expectation become available as a consequence of the mere occurrence of the stimuli. They have been formed in response to the animal’s experience over time in observing sequences of stimulus constellations. Think of a two-layer neural network whose connection weights from its input layer to its output layer correspond to the adjustable associative strengths. In response to input patterns the network computes composite expectations in the form of the activity levels of the output units. Target output values representing USs, provided by so-called “teaching inputs,” are compared to the network’s actual outputs—the surprise computation—to determine the error that drives learning. In addition to participating in this comparison, these expectations also directly determine the strength of the animal’s tendency to produce a CR.

This process does not require a scanning of the organism’s memory for previously experienced instances of the stimulus constellation that is currently present: this experience has been cached in the connection weights, and the network reads out an expectation in response to the current input pattern. In a neural network setting that considers the relative timing of inputs (i.e., the teaching input is whatever stimulus pattern occurs shortly *after* the input pattern setting the activation levels of the input units), the network becomes a *predictor*, meaning that each of its output patterns will tend to resemble the input pattern that comes next. (Of course this assumes the network is complex enough to represent the prediction function.) The process is not tied to a specific US. The network’s weights summarize, in a statistical sense, the totality of the organism’s previous experience as to what

stimulus constellations tend to follow other stimulus constellations. In machine learning, one would say that a *forward model* of environmental contingencies is learned via *supervised learning* (Barto, 1990).

Other concepts have been proposed for how an expectation for associative learning might be implemented in the nervous system. For example, Grossberg (1982) proposed that an expectation is a feedback pattern of neural activity derived from signaling across an entire network gated by long-term memory, and that unexpected events trigger a “mismatch-modulated arousal burst,” i.e., what we would call a surprise signal.

2.2. BAYESIAN SURPRISE

A formal theory of surprise was proposed by Itti and Baldi based on the Bayesian framework (Itti and Baldi, 2005, 2006, 2009). In this framework, probabilities, which correspond to subjective beliefs, are updated as new observations are made using Bayes’ theorem to convert prior beliefs into posterior beliefs. What they call *Bayesian surprise* is a measure of the difference between an observer’s prior and posterior beliefs.

Here is how they formalize this. An observer is assumed to have background beliefs characterized by a prior probability distribution over hypotheses or models of its world, M , that are in some space of models, \mathcal{M} :

$$\{P(M)\}_{M \in \mathcal{M}}.$$

Upon obtaining new data D , the observer updates this prior distribution into the posterior distribution by applying Bayes’ theorem:

$$\forall M \in \mathcal{M}, \quad P(M|D) = \frac{P(D|M)}{P(D)} P(M).$$

Bayesian Surprise is a measure of the dissimilarity between the prior and posterior distributions. Itti and Baldi do this using the relative entropy, or Kullback-Leibler (KL) divergence, between these distributions:

$$\begin{aligned} S(D, \mathcal{M}) &= KL(P(M), P(M|D)) \\ &= \int_{M \in \mathcal{M}} P(M|D) \log \frac{P(M|D)}{P(M)} dM. \end{aligned}$$

This measure gives the amount of information needed to transform the prior into the posterior distribution:

A unit of surprise—a “wow”—may then be defined for a single model M as the amount of surprise corresponding to a two-fold variation between $P(M|D)$ and $P(M)$, i.e., as $\log P(M|D)/P(M)$ (with log taken in base 2), with the total number of wows experienced for all models obtained through the integration [in the equation above]. (Itti and Baldi, 2009)

According to this theory, surprise is a measure of the discrepancy between beliefs before and after an observation. A surprising event is one that is not well predicted by the animal’s current beliefs formed in response to its previous experience. In this case, the expectation that determines surprise is the

set of beliefs held by the agent before the observation in question, that is, the prior probability distribution over possible world models: $\{P(M)\}_{M \in \mathcal{M}}$. Itti and Baldi (2005, 2006, 2009) argue that this definition has key advantages over alternatives in being more principled, more widely applicable, and more able to account for what attracts human visual attention. Importantly for our purposes, these authors also discuss how assessing surprise differs from detecting statistical outliers, which is one of the notions commonly (though erroneously we will argue) invoked for detecting novelty. We discuss this in Section 5 where we examine differences between surprise and novelty.

Schmidhuber and colleagues (Schmidhuber et al., 1994; Storck et al., 1995) proposed using Bayesian surprise (as later defined by Itti and Baldi) as a measure of learning progress for reinforcement learning agents. This measure of surprise generates a “curiosity reward” that encourages the agent to behave so as to continue learning efficiently by seeking regions of its environment where it is surprised while avoiding regions where it is “bored,” either because it has already learned as much as it can in those regions (thereby eliminating surprise) or because there are no learnable regularities (so that surprise is absent because new information is not acquired). This is one of the first proposals for how ideas related to what psychologists call intrinsic motivation can be implemented in a machine learning system, and much additional research has been lately done in this area (Baldassarre and Mirolli, 2013).

Itti and Baldi (2005, 2006, 2009) were concerned with attention rather than learning, but their concept of surprise arises from the Bayesian approach to learning where a prior belief distribution is updated by Bayes’ theorem to a posterior distribution upon each new observation. Large Bayesian surprise means that learning from a new observation has made a large change in the animal’s beliefs about the contingencies in its world. In its most general form, Bayesian learning does not explicitly involve the computation of prediction errors. Instead of processing errors generated by an existing model, learning processes evidence for all possible models and updates beliefs accordingly. Unlike error-correction learning, where the error as a measure of surprise is the direct driving force of learning, Bayesian surprise is the result of learning but not its direct cause, coming after the Bayesian update instead of before it.

However, Bayesian learning can be approximated, and in some cases computed exactly, by an error-correction process. The Kalman filter, for example, uses error correction to perform Bayesian learning in the context of linear-Gaussian systems (Welch and Bishop, 1995). The mean of the Gaussian posterior distribution is updated by multiplying the innovation, or prediction error, by the Kalman gain which controls the allocation of weight between the prediction of a current model and a new observation based on a measure of confidence in the model and in the observation. Bayesian learning can be approximated in a number of ways, such as through the Laplace approximation and variational methods (Bishop, 2006), that permit updates to be made on the basis of prediction errors. Variational approximation plays a key role in the hierarchical architecture proposed by Mathys et al. (2011), who discuss the relationship of the

resulting learning process to error-correction methods like the Rescorla-Wagner model.

Models of classical conditioning based on Bayesian methods, including the Kalman filter, have been proposed that go beyond the account provided by the Rescorla-Wagner model (Dayan et al., 2000; Kakade and Dayan, 2002a; Courville et al., 2004, 2006). Changes in the world, and therefore changes in the correct world model, are sources of Bayesian surprise. Bayesian methods not only update beliefs in specific models but also the confidence in those beliefs, and surprise causes *decreased confidence* in current beliefs. As a result, new observations should be given more weight than previous observations (as in the Kalman filter), implying that the speed of learning about the uncertain predictive relationships should increase. This Bayesian account of increases in the rate of animal learning observed in certain experiments (Rescorla, 1971) accomplishes what the Pearce-Hall model (Pearce and Hall, 1980) does via its use of an explicit measure of surprise as the magnitude of a prediction error. TD learning has also been developed in a Bayesian framework (Engle et al., 2003).

Another area in which prediction errors appear in a Bayesian framework is in the “predictive coding” architectures of Rao and Ballard (1999) and Friston and Kiebel (2009). These are layered hierarchical systems going from input levels to levels encoding information in a more abstract fashion. The key aspect of these systems is that the bottom-up flow of information from sensations to abstract representations is paralleled by a top-down information flow where the top levels project predictions to the lower-levels. This allows higher-level stages to receive information only through the information mismatch between their predictions and sensations, so that higher levels receive only unpredicted information. Prediction errors are used to propagate information from the bottom up to the higher levels of the system, and also to continuously update the top-down predictors. These proposals refine the concept of surprise as they capture surprise at multiple levels, namely from the prediction of simple, isolated events at the lower levels, to the prediction of the behavior of more complex compounds of items at the higher levels.

2.3. INFORMATION THEORETIC SURPRISE

Although Itti and Baldi’s Bayesian theory of surprise is connected to information theory (KL divergence is a measure of information gain), other concepts of surprise are more directly based on information theory. One example is what Tribus (1961) called *surprisal* to refer to the self-information of the outcome of a random variable, which is a measure of the information content of the outcome. If outcome ω occurs with probability $P(\omega)$, then the self-information, or surprisal, is $-\log P(\omega)$. Thus, an outcome that is highly unlikely has high surprisal when it occurs. The expected value of surprisal for observations drawn from a random source is the entropy of that source. Computational linguists, e.g., Roark (2011) and Monsalve et al. (2012), use the term *lexical surprisal* to refer to the negative log of the *conditional probability* of a word in a sentence given the preceding words in the sentence. Although Tribus’ definition of surprisal does not explicitly invoke conditional probabilities, there is always an implicit assumption that surprisal is conditioned on a context or model. Therefore,

when we refer to surprisal below, we always have a conditional form of surprisal in mind.

An important contrast can be drawn between surprisal and Bayesian surprise. The usual example is to consider viewing a television screen showing white noise, or “snow” (Schmidhuber et al., 1994; Storck et al., 1995; Itti and Baldi, 2005). After a while this becomes very boring even though the information content of each frame, or its surprisal, is very high because there are so many equally-likely patterns of random noise. On the other hand, a viewer’s Bayesian surprise will decrease and eventually disappear as their beliefs adjust so that random frames become expected. “Thus, more informative data may not always be more important, interesting, worthy of attention, or surprising” (Itti and Baldi, 2005).

Tribus’ notion of surprisal plays a prominent role in the global brain theory of K. Friston and colleagues which is based on the principle of “free-energy minimization” (Friston et al., 2006; Friston, 2009, 2010). This principle states that intelligent agents aim to minimize a free energy function of their internal states. If one assumes that an agent maintains a model of the causes of its sensory input, this principle implies that intelligent agents act on their environments to avoid surprises, which means working to make observations that conform to their expectations. Another component of this theory is that intelligent agents learn by revising their models to make more accurate predictions. These implications can be seen to follow from free-energy minimization through the perspective of variational Bayesian inference. Free energy (in this case the variational free energy) is always greater than or equal to the negative log of the evidence, or the marginal likelihood, of the agent’s model. Model evidence is the probability of observations given the agent’s current model: if s denotes an agent’s sensory state at some time and M denotes its current model, the model evidence is $P(s|M)$ (where hidden states have been marginalized out). Thus, acting to minimize this free energy function tends also to minimize the negative log of model evidence (since the latter quantity is always less than or equal to the free energy). This is equivalent to tending to maximize the (positive) log of model evidence, which is the same as tending to maximize the model evidence itself since the logarithm is a monotonically increasing function. The theory’s connection to surprisal is due to the fact that the negative log evidence for a model is the surprisal conditioned on that model, $-\log P(s|M)$, so that maximizing model evidence is the same as minimizing this notion of surprise. According to this theory, then, intelligent agents act on their environments to suppress discrepancies between their predictions and what they actually experience, that is to avoid being surprised.

The theory also relates to Itti and Baldi’s (and Schmidhuber’s) notion of Bayesian surprise. In addition to acting to increase evidence of a current model, agents can reduce free energy by adjusting their model to make more accurate predictions. Through a learning process, a current probability distribution over models (a prior distribution) is updated to a new distribution (a posterior distribution) that takes into account each new observation. As the model becomes more accurate, the KL divergence between these distributions—that is, the Bayesian surprise—decreases, which decreases free energy. The Bayesian surprise becomes zero only

when the model makes perfect predictions. An additional implication of this theory arises from the role of model evidence in Bayesian model comparison, where there is an automatic penalty for model complexity. This implies that the work done by agents to increase how well their model accounts for observations is balanced by a tendency to minimize model complexity, a form of Occam's razor. Friston and colleagues present hypotheses about how the brain might implement the elements of this theory (Friston et al., 2006; Friston, 2009, 2010).

In his book "Novelty, Information, and Surprise" Palm (2012) provides definitions of all three of these terms. Roughly, novelty is the same as Tribus' surprisal, but surprise is given an interesting definition that depends on the concept of a "description," which is a mapping from possible outcomes of a random variable to propositions that are true for a collection of outcomes. A key aspect of this theory seems to be that by knowing the description an observer is using, that is, by knowing the whole mapping, it is possible to consider the probability that an outcome will have the same description as the outcome observed. Then the amount of surprise experienced by an observer depends not on the probability of the observation, but on the probability of any observation with the same description. Palm gives the following example. Suppose that in a state lottery the sequence of numbers (1, 2, 3, 4, 5, 6) were to be drawn. This would be much more surprising than the sequence (5, 11, 19, 26, 34, 41) even though both sequences have the same probability of being drawn. "The reason for our surprise in the first case seems to be that this sequence can be exactly described in a very simple way: it consists of the first six numbers. . . . it is much more probable to obtain a sequence that does not admit a simple exact description In the special case of (1, 2, 3, 4, 5, 6) we could argue that there are only two such extremely simple sequences, namely the last 6 and the first 6 numbers" (Palm, 2012, p. xix). Palm argues that his extension of classical information theory allows one to incorporate a "person's interests, intentions, and purposes." How this intriguing view of surprise relates to the more familiar ones discussed above is not yet completely clear to the authors.

2.4. SUMMARY

According to the commonsense notion as well as the most prominent formulations, surprise involves a comparison between an expected and an actual observation. The comparison does not need to entail a scan of the contents of memory. Expectations formed on the basis of past experience can be linked directly to stimuli so that they are aroused by the occurrence of those stimuli, or aroused by an inference process in the absence of those stimuli. Surprise is a measure of the discrepancy this comparison reveals, whether it is a simple signed difference as in error-correction learning rules, the KL divergence in Itti and Baldi's Bayesian surprise, or some other measure. Predictive coding by hierarchical systems suggests how surprise might be generated at different levels of abstraction. The term surprisal has been proposed for an observation's self information, a quantity inversely related to the probability of the observation conditional on a model. Bayesian surprise and surprisal differ in significant ways. Friston's global brain theory based on the free-energy principle suggests that intelligent agents act in order to reduce surprisal conditioned on

their current models, while they also reduce (future) Bayesian surprise by adjusting their models to make better predictions.

3. NOVELTY

Confronting the problematic concept of novelty, Berlyne (1960) emphasized a number of relevant distinctions. First, he distinguished between *short-term*, *long-term*, and *complete* novelty. Something may never have been encountered before (complete novelty), or not encountered in the last few minutes (short-term novelty), or not encountered for some intermediate time, e.g., a few days (long-term novelty). Another distinction is that between *absolute* and *relative* novelty. A stimulus is *absolutely novel* when some of its features have never been experienced before, whereas it is *relatively novel* if it has familiar features but they occur in some combination or arrangement that has not been previously encountered.

Berlyne claimed the following:

Any new experience, even if it does not seem to be a combination of familiar experiences, must have some definite degree of resemblance to experiences that have occurred before. It will inevitably be possible to insert it into an ordering of familiar stimuli or to assign to it values among dimensions that are used to classify them. (Berlyne, 1960, p. 19)

He gives the example of seeing a man taller than any seen before: it is still possible to place the experience on a familiar scale, or more generally, to locate the experience in the appropriate multidimensional feature space. Further, according to Berlyne:

For any adult human being, or even any adult dog, cat, or rat, a new stimulus must be similar to, and relatable to, a host of familiar and frequently experienced entities. However, bizarre a non-sense figure may be that is shown to a human adult, it must consist of lines, angles, and curves such as he has seen on countless occasions. (Berlyne, 1960, p. 20)

Note that Berlyne restricts this comment to adults. The situation must be different for young children, due not only to their relative lack of experience but also due to the deeper need to establish the feature spaces and dimensions that are useful for categorizing experience. For designers of artificial agents this is a key issue.

Berlyne's distinctions are important because they connect to our ordinary understanding of what the term novelty means while revealing some of the issues that make the concept problematic. In formal notions of novelty to which we now turn, the links to our commonsense notion are not always apparent.

3.1. MEMORY-BASED NOVELTY

The simplest translation of our commonsense idea of novelty into a more precise notion is that the novelty of an event is assessed by examining a memory store of past observations, where a memory system might require more than one experience of an event to form a lasting memory. An observation is completely novel, to use Berlyne's term, if a representation of it is not found in memory. If memory fades with time, this process assesses short-term or long-term novelty depending on the fading rate. This of course ignores many aspects both of novelty and of memory, and it may not be feasible from a computational perspective.

But some more sophisticated methods for novelty detection are elaborations of this basic idea. Novelty detection based on clustering is one example. Using a distance measure based on similarity, data can be clustered into classes so that items in a class are “close” to one another and not close to items in the other clusters. Novelty here means that an item is not close enough to the mean of an already existing cluster, so that a new cluster needs to be formed. There are very many clustering methods, and there are many methods for determining when a new cluster should be added (Markou and Singh, 2003).

Determining distances from existing clusters is a search of a memory that stores the cluster means, making it more feasible than a naive memory-based method. Prominent neural network methods for novelty detection, such as methods based on self-organizing feature maps (Kohonen, 1984; Nehmzow et al., 2013), perform this basic process where the memory scan is performed in parallel by the network. Of current interest in statistics and machine learning are Bayesian non-parametric clustering methods (Gershman and Blei, 2012). Instead of specifying the number of clusters in advance, these methods allow the number of clusters to grow as new data items arrive. These methods do not involve a literal scan of memory, but determining whether a new cluster is needed essentially relies on determining that none of the existing clusters properly explains the data.

Another kind of memory-based novelty arises in the case of content-addressable associative memory systems. Perhaps the most well-known and simplest is the correlation matrix memory proposed by Kohonen (1977, 1980, 1984). Instead of being stored in separate memory locations, information is superimposed and distributed across a memory substrate, for example a neural network, and retrieval is a kind of filtering process. The stored items are vectors of real numbers, and the memory is a matrix formed from the stored vectors in such a way that upon being presented with an input vector, the system produces as output a weighted sum of all the stored vectors, where each weight is a measure of how well that stored vector correlates with the input vector. When the input vector is a distorted version or a fragment of a stored vector, it is expected that it will correlate most strongly with that vector and much less with the other stored vectors, implying that the memory’s output will be a less noisy version of the input vector or a “completion” of it. Mathematically, the memory’s output is the orthogonal projection of the input, x , onto the subspace, \mathcal{L} , spanned by the stored vectors, which is the vector, call it $\hat{x} \in \mathcal{L}$ that is “closest” to x . Every vector x can be expressed as the sum of \hat{x} and a vector, \tilde{x} , in the subspace orthogonal to \mathcal{L} . Kohonen (1977) says that “ \tilde{x} is the amount that is ‘maximally new’ in x . It may be justified to call this component the ‘novelty,’ and the name Novelty Filter is hereupon used for a system which extracts \tilde{x} from input data x ...”. Roughly, then, this kind of novelty refers to those fragments or aspects of an observation that are not fragments or aspects of previously stored experiences. Our memory systems are undoubtedly much more complicated than a correlation matrix memory, but it is worth keeping this example in mind when we discuss associative novelty as studied in neuroscience in Section 6.2 below.

3.2. NOVELTY AS STATISTICAL OUTLIER

A common notion is that an observation is novel if it is a *statistical outlier*, meaning that it is significantly different from other members of the sample from which it is drawn. In general terms, detecting outliers requires modeling the usual distribution of observations and detecting when an observation departs significantly from the model. Sometimes this is called *anomaly detection*. Many methods have been proposed to detect outliers and to handle them, but what concerns us here is what being an outlier means with respect to our common idea of novelty and how it differs from surprise.

One area in which this idea of novelty plays a prominent role is machine learning. For example, learning a classification rule by supervised learning involves adjusting a classifier’s parameters on the basis of training examples drawn from a corpus of labeled examples. It is important that the corpus of training examples is representative of the input data to which the classifier will be applied. Novelty detection for supervised learning is the problem of determining if an input does not belong to the class of inputs represented by the training examples, i.e., determining if the input is an outlier. For novel inputs, the output of the classifier will be considered unreliable.

Nearly all the statistical approaches to this problem model the probability density of the training data and identify inputs as novel if they fall in regions of low estimated density. Many methods exist for estimating probability densities from a finite number of samples, both parametric or non-parametric (Duda and Hart, 1973; Markou and Singh, 2003), and many methods have been suggested for how to use the estimated probabilities to determine when an input should be regarded as novel. The details of these methods need not concern us here; the principle remains the same: according to this view *novelty means having a low estimated probability of occurrence*. Note that according to the definition of surprisal given in Section 2, this is the same as saying that being novel means having high surprisal, a point to which we return in Section 5 below.

We commented in Section 2.3 that although Tribus’ definition of surprisal does not explicitly invoke conditional probabilities, there is always an implicit assumption that surprisal is conditioned on a model or context. Estimated probabilities for outlier detection are conditioned on the context of the collection of samples and background assumptions about the sample space. This raises questions about equating novelty with “low probability” because it is based on the assumption that the system can represent the entire domain of possible samples in advance of experiencing them, and so can assign zero probability to all instances not observed up to a given moment. An aspect of our commonsense notion of novelty for which this view is not able to account is the possibility that an observation might occur that the system is not able to represent in terms of existing categories. Assuming that the sample space consists of all possible configurations of the lowest-level sensor readings may be a solution for artificial systems (e.g., the pixels of a camera), but it seems an inadequate account of biological memory which is typically not so eidetic. Indeed, as we discuss in Section 6 below, novelty may trigger brain activity whose function is to acquire new representations.

3.3. SUMMARY

Berlyne (1960) distinguished between several difference senses in which the term novelty is used, and formalizations of novelty are not as unified as those of surprise. Straightforward interpretations involving searches of memory for previous encounters do not do justice to the complexity of either the concept of novelty or of the nature of memory. Clustering-based concepts expand naive memory search and make better contact with the commonsense notion of novelty as the quality of being different from what is in a memory store. Content-addressable associative memory systems suggest a more abstract notion of novelty as, roughly, fragments or aspects of an observation that were not present in previous experiences. Statistical interpretations in terms of outlier detection have many applications, but as we argue below they also abstract away from important aspects of our commonsense understanding. In neuroscience additional categories of novelty are described, which we discuss in Section 6.

4. NOVELTY AND SURPRISE: TYPICAL FEATURES

We have seen that there are various proposals about how to define surprise and novelty, all having some strengths. On this basis, we think it is premature to propose definitive definitions. Nevertheless, we also think it is possible and useful to highlight the main features of the two concepts that represent the “poles” around which the different definitions should gravitate. **Table 1** displays these features, and we now briefly explain them.

A key difference between novelty and surprise is due to the type of knowledge store they use and the way they process such

knowledge. Novelty is based on memory stores and the processes that determine if a given item is, or is not, in the store. Surprise, on the other hand, is based on expectations of systems capable of predicting, the processes generating such expectations, and the processes that compare the expectations with what is actually experienced. An observation is novel when a representation of it is not found in memory, or, more realistically, when it is not “close enough” to any representation found in memory. Novelty triggers the formation of new representations for entry into long-term memory. These representations can then be exploited to perform other cognitive processes, including the generation of surprise by exploiting already existing representations (Lisman and Grace, 2005; Kumaran and Maguire, 2007). The case of surprise is different because its core element is not the incoming item but the predicted item. Indeed, the incoming item can be either familiar or novel—this does not count. What counts for surprise is that the system perceives “something” that is different from the prediction, whatever that “something” is.

Novelty and surprise also differ with respect to their relation to time. The expectations or predictions that underly surprise have to do with the dynamic flow of events happening in time (with the possible exception of spatial predictions underlying Berlyne’s notion of incongruity, which may, however, involve the visual scan of a stimulus, thereby again involving time). Predictions typically involve a specific time, or range of times, in the future when something is expected to happen: “If I see A at time t , then I expect to see B at time t plus something.” Novelty, on the other hand, seems not to be strictly related to time. The

Table 1 | The typical features of novelty and surprise.

Features	Novelty	Surprise
Type of knowledge store, process involved	Memory, memory recall	Predictor, prediction
Variants of the knowledge and process involved	- Formation of new representations - Formation of new links between the representations of the features/components of the novel data	- Deterministic expectations - Stochastic expectations
Time	Time not a key factor: items in memory are always available for comparison	Incoming data usually compared with a temporalized prediction
Processes for novelty/surprise triggering	One phase: - Experience does not match memory	Two phases: - Formulation of prediction - Prediction is violated
Typical functions	- Support the formation of new representations - Generate learning signals for the sub-component detecting novelty, or for other sub-components - Direct/motivate attention and learning resources to novel stimuli	- Support the improvement of predictions - Generate learning signals for the predicting sub-component or for other sub-components - Direct/motivate attention and learning resources to unpredicted stimuli

comparison of current experience with the contents of memory, i.e., the process that supports novelty detection, is not sensitive to the time at which a memory was formed, nor to the time the novel item is perceived: what really matters is only the absence of a representation of the perceived stimulus in memory. Berlyne's distinction between short-term, long-term, and complete novelty refers to differences in how this process may work, but in none of these cases is the timing of the perception as critical as it is for surprise.

Both surprise and novelty increase an animal's level of arousal, direct its attention, enhance learning, and elicit other appropriate behavior. But in some other respects surprise and novelty differ in their typical functions. Where novelty often supports the acquisition of representations, surprise supports the improvement of predictions. More specifically, novelty supports the acquisition of items by memory, while surprise plays a key role in improving the capacity of the system to predict (as in error-correction learning reviewed in Section 2.1) or to signal that such an improvement has taken place (as in the Bayesian account as discussed in Section 2.2).

5. RELATIONSHIP BETWEEN SURPRISE AND NOVELTY

Surprise often—perhaps always—accompanies novelty, which may be a major reason the two concepts tend to be confounded. Indeed, if one assumes that an agent is always making predictions about what it is going to soon experience, encountering something novel should not only trigger a novelty response, because no representation has been found in memory that corresponds to the perception, but also surprise, because the agent's expectations must be violated by the novel item which could not have been predicted. In this case, the agent is not predicting that it will not observe that item, but it is predicting that it will observe something else—a prediction that is violated. Whether or not this argument is convincing depends upon whether animals are always expecting something, which in turn depends on what it really means to expect something, which we will discuss shortly.

On the other hand, it is clear that surprise does not imply novelty. A familiar observation may be surprising in a context in which something else is expected. It is easy to come up with examples: for instance, we can be surprised at finding our car door locked when we thought we had just clicked the unlock button on the key fob.

A more interesting example is provided in a study by Huron (2004) of laughter in listeners to Peter Schickele's PDQ Bach compositions. In this example, the expected "something else" is in fact rare, whereas the actual observation is familiar, though unexpected. Schickele has composed a large number of humorous pieces attributed to the fictional P.D.Q. Bach. Huron argues that a plausible explanation for the laughter these compositions induce is that laughter occurs at "dramatic violations of expectation." In one composition (*Quodlibet for Small Orchestra*), Schickele reproduces a well-known theme from a Beethoven symphony, but instead of continuing with Beethoven's finish to the movement "which is the rarest continuation in Western music with a probability of less than 0.007," he switches to a "musically banal" conclusion. Invariably, listeners burst

into laughter at the moment of this switch. Huron (2004) summarizes:

In short, Schickele's transgression here is a violation of veridical expectation ("That's not how the music goes.") rather than a schematic transgression ("That's not what happens in music.") The violation is amplified by the extreme contrast between veridical and schematic probabilities. (Huron, 2004, p. 702)

What Huron means by a "veridical expectation" is an expectation created through past experience with the specific music in question, in this case Beethoven's symphony, which—during listening—generates an expectation for its usual ending. But the usual ending is rare in music in general, that is, its probability of being heard is very low, whereas Schickele's ending has much higher probability. Therefore, the "schematic transgression" is a mismatch between an expectation for something unlikely and the receipt of something familiar.

As discussed above in Section 3.2, a common formalization of novelty in machine learning is that being novel means being a statistical outlier, and novelty detection is accomplished by modeling the probability density function of possible observations and regarding an observation as novel if it falls in a region of low enough estimated density (according to a given threshold or a more sophisticated criterion). We are not aware of claims that this formalization of novelty provides a good account of what novelty means for an animal, but it is pertinent to ask if this notion of novelty is consistent with either our common-sense understanding of the term or novelty's typical features. The answer has to be no. It is true that if the probability of an event occurring is low, then the probability that a representation of that event is stored in memory is low as well. But it is clearly missing something important about novelty to equate low estimated probability of occurrence with novelty. It is easy to think of examples of events that are not novel at all but that have a very low probability of occurring. For example, any event that occurred only once in the past, and that is distinctly different from other experienced events, will likely be assigned a low probability of occurring again. But that event may be vividly memorable and therefore familiar if it were to happen again. Furthermore, if so-called novelty detection happens as a result of a mismatch between one's estimated probabilities and current perceptions, this seems to be a clear case of surprise rather than novelty, as discussed in Section 2. Thus, while treating low probability events as novel may be a good method for machine learning, it is a poor model of what novelty really is and represents a misleading use of the term.

The same reasoning explains why Tribus's term surprisal (Tribus, 1961) is more consistent with what we mean by surprise. Indeed, the surprisal value of an observation, that is, a measure inversely related to its probability of occurring, can be thought of as the discrepancy between its probability of occurring and the fact that it actually occurred. Thus, surprisal appears to be consistent with the notion of surprise according to our analysis (despite the fact that it is basically the same as novelty according to the statistical outlier view of novelty). Surprisal is particularly consistent

with our characterization of surprise when it is explicitly conditioned on a context as in the lexical surprisal of computational linguists (Monsalve et al., 2012; Roark, 2011). In this case, surprise as surprisal is triggered by an event occurring in a context in which the estimated probability of its occurrence is low.

Itti and Baldi's (2005, 2006, 2009) Bayesian surprise is not a misleading use of the term since their definition is based on a discrepancy between beliefs before and after an observation. The degree of surprise generated by an observation depends on how strongly it changes the probability distribution over models that characterize an observer's beliefs about how its world works. It is not clear that the Itti/Baldi notion is the only, or the best, Bayesian account of surprise, but this account of surprise is consistent with what we regard as its typical features.

Bayesian surprise has interesting implications with respect to the view of surprise as surprisal. Here is a slightly modified version of an example given by Itti and Baldi. Consider incoming data, D , that has a very low probability given the current context C , that is, D is surprising in the sense of having high surprisal. Suppose the observer has only two models, and the observation has a low probability given the context and either model, that is, $P(D|C, M_1)$ and $P(D|C, M_2)$ are both low. In this case, even though the surprisal of D is high, Bayesian surprise would be very low since D has little effect on the agent's beliefs: it is not useful in discriminating between M_1 and M_2 . This is a very hypothetical example, but it raises the question of which account of surprise is more consistent with the processes that generate surprise in animals.

6. SURPRISE AND NOVELTY IN NEUROSCIENCE AND COGNITION

This section considers some important threads of neuroscience research related to surprise and novelty. Enlisting the concepts developed in the previous sections shows how existing results might be reinterpreted in a way that improves our understanding of behavior and the neural machinery underlying it. The goal here is not to cover the large neuroscience literature related to novelty and surprise, but rather to show how keeping the distinction in mind may be a useful heuristic for isolating interesting problems and seeking answers to questions about how surprise and novelty are processed in the brain. Thus, below we focus on a selection of biological cases that involve mechanisms where the distinction between novelty and surprise is blurred or controversial, while omitting consideration of other brain phenomena more reliably associated to each of the two concepts (e.g., cerebellum, forward models, prediction errors, classical conditioning; anterior cingulate cortex, anticipations, error-related negativity; amygdala, classical conditioning).

Modern neuroscience literature distinguishes between three types of novelty to which the brain responds: stimulus novelty, contextual novelty, and associative novelty (Ranganath and Rainer, 2003; Kumaran and Maguire, 2007). These three types of novelty are investigated with different experimental paradigms, involve partially overlapping networks of brain areas, and are based on various neural mechanisms. In addition, an important thread of neuroscience research deals with what have been called dopamine "novelty responses." In what follows we discuss these

four novelty categories in turn, trying to clarify whether the term "novelty" is an appropriate label or if the investigated phenomena have more to do with surprise.

6.1. STIMULUS NOVELTY

Stimulus novelty refers to the phenomenon for which the neural and behavioral responses to a particular stimulus (e.g., the sight of an object) change when it is experienced multiple times. A typical observation is that with repetition of a stimulus the neurons responding to it present a progressively decreasing activation, a phenomenon called *repetition suppression* (Ringo, 1996; Henson and Rugg, 2003). Repetition suppression is stimulus specific and has been observed in various types of experiments, from classification (Sobotka and Ringo, 1994) to delayed-matching-to-sample tests (Li et al., 1993). Some of the areas most sensitive to the novelty of stimuli are inferotemporal cortex (Ranganath and Rainer, 2003), an area involved in object recognition, the perirhinal cortex (Brown and Aggleton, 2001), an area close to the hippocampus and involved in episodic memory, and the prefrontal cortex (Asaad et al., 1998), the highest multimodal associative cortex.

Stimulus novelty seems to be the classical case of novelty, where the incoming items trigger novelty detection when they do not correspond to any existing memory. The novel items trigger the formation of a neural representation at multiple levels within the brain areas mentioned above, so they progressively became familiar (Ranganath and Rainer, 2003).

An intriguing issue related to stimulus novelty arises from the fact that novel items seem to cause an initial high activation of the brain areas where novelty is presumed to be computed. This raises a twofold question: (a) what are the specific mechanisms that cause such a high activity, and (b) what is its adaptive function? While the question about mechanisms is an interesting challenge for computational modeling, the view that the main function of novelty detection is the formation of representations of the novel items in memory might explain why novel items cause higher activation. Learning often needs to be supported by the production of neuromodulators. The elevated activation caused by novel items might trigger the production of neuromodulators, for example, noradrenaline and acetylcholine (see Ranganath and Rainer, 2003, for a review). In turn, the presence of neuromodulators may support the formation of new neural representations. This hypothesis suggests a number of neuroscientific investigations directed toward understanding the brain mechanisms implementing the various steps of the suggested causal chain, as has already happened with respect to dopamine and hippocampus, which are involved in the other types of novelty detection considered below (see Lisman and Grace, 2005).

6.2. ASSOCIATIVE NOVELTY

Associative novelty is one of the most subtle and interesting cases of novelty studied in neuroscience. Associative novelty refers to situations where familiar stimuli are associated in novel configurations (Kumaran and Maguire, 2007). The associations can be: *spatial*, where familiar items appear in new spatial locations; *item-item*, where items appear in novel combinations, e.g., two familiar words are paired in an odd fashion; or *temporal*, where familiar

items appear in a novel temporal sequence. Interestingly, the field of associative memory is one in which the blurring of the distinction between novelty and surprise is most prevalent. An example is given by the following from O'Keefe and Nadel (1978) with our italics:

Imagine that you are in a classroom . . . *suddenly*, your attention is diverted when a naked man enters the room. . . . the entrance of the naked guy was a *novel event* in that it was *unexpected* and *out of context*. . . . *novel events* attract attention and they are more effectively *encoded in memory* than are *predictable events*.

Associative novelty includes cases that are most difficult to classify, including some that may involve *both* novelty and surprise.

Temporal associative novelty involves a paradigmatic case of surprise: if you perceive a familiar item in a novel temporal sequence, it seems that items that precede the target item constitute the context that supports an expectation which is violated by the appearance of the familiar target item. Hence surprise.

The spatial case is also probably related more to surprise than to novelty. When we perceive a familiar item in a new spatial location, we already have its representation in memory. It is likely that finding the item in a position where we never experienced it just violates our expectation regarding its position—hence surprise. This interpretation is consistent with the fact that in experiments dealing with spatial associative novelty, subjects are typically exposed to the associative pairings for many times before their familiarity/novelty discrimination responses are assessed (Duzel et al., 2003; Kohler et al., 2005). It is most likely that these repeated exposures are needed for expectations to be created, so that they can be violated to trigger the inappropriately-labeled “novelty” signal.

Item-item associative novelty seems to be the more complicated case to classify. To understand whether a case is best called novelty or surprise may require knowing which brain mechanisms are involved. It is well accepted that the hippocampal system is involved in the formation of complex episodic memories and seems to play a critical role for the detection of multiple kinds of novel associations (Wan et al., 1999; Brown and Aggleton, 2001). The *comparator hypothesis* is one of the most established hypotheses about how the hippocampal system detects associative novelty. It refers to the following processes (Hasselmo and Schnell, 1994; Kumaran and Maguire, 2007; Duncan et al., 2012): (a) familiar aspects of the percept (“lures”) actively recall previous memories on the basis of pattern-completion-like mechanisms, for example, an item recalls other items previously experienced in association with it, and (b) some of the perceived items mismatch with the recalled items so that a mismatch signal is triggered. If this theory is correct, than associative novelty is closely related to Berlyne's notion of incongruity, which we classified as a form of surprise in Section 2 because it involves a mismatch between explicit expectations/predictions and incoming data. Kohonen's “novelty filter” (Kohonen, 1977) described in Section 3.1 is relevant to this point: the novelty in an input is, roughly, that part of it that is not predicted by the remaining part. However, it might also be the case that sometimes sets of items are grouped into single compound representations, and that the brain, by searching in

memory for these representations and not finding any, registers observation of the set as actual novelty. It is also plausible that in such circumstances both novelty and surprise are simultaneously at play.

The general point here is that some areas of the brain, especially higher-level associative areas such as the hippocampus, may use the same machinery to exploit the representations of associated items to either detect novelty or to detect surprise, depending on the context and the task at hand, and that in some cases both novelty and surprise may be registered. What are the actual mechanisms that the brain uses in each circumstance is an important question for neuroscience research.

6.3. CONTEXTUAL NOVELTY

Contextual novelty is another type of widely-studied novelty, closely related to associative novelty (Ranganath and Rainer, 2003). This refers to the behavioral and neural reactions to stimuli that are familiar but are unexpected given the context in which they occur. Contextual novelty is often studied in *oddball* experiments where, for example, sequences of a repeating auditory stimulus (e.g., a simple tone) are interleaved with rare odd signals (e.g., a “moo” of a cow) (Ranganath and Paller, 1999). The reaction of the brain to an oddball stimulus is often monitored via electric field potentials (Event-Related Potentials—ERP) generated when the brain detects the stimulus. The typical result of these tests is the manifestation of a positive wave of the electric field happening about 200–300 ms after the odd stimulus and named “P300” or “P3” (Friedman et al., 2001). Intense investigation has led to the isolation of a P3a component of the wave, also called “novelty P3” (Soltani and Knight, 2000). Various studies indicate that the novelty P3 originates from a network of brain areas including the hippocampal system considered above (Soltani and Knight, 2000). This and other elements suggest that overlapping brain machinery might underline associative novelty and contextual novelty (Kumaran and Maguire, 2007).

It is easy to see that in the case of contextual novelty the mechanisms of prediction and surprise, and not of novelty, are in action. Indeed, in the oddball experiments the odd item is appealed to as “novel” even if it is often a familiar item that is presented to the participants in an unpredictable fashion, e.g., a “cow moo” presented after a sequence of simple tones. In this case, the “moo” is surely not novel as the participants have surely heard that sound several times before the experiment. Instead, the “moo” represents a typical example of familiar item that generates surprise because it is unpredicted after a sequence of regular tones. We expect that the clear recognition of what phenomenon is being observed, in this case surprise, will help researchers to recognize new problems and new solutions to them, and to suggest experiments that will lead to a better understanding of the brain processes involved.

6.4. DOPAMINE “NOVELTY” RESPONSES

Another important example of the confusion between surprise and novelty can be found in the recent neuroscience literature on dopamine. Dopamine is a neuromodulator that is well known to play a pivotal role in motivational and reinforcement learning processes (Wise, 2004; Berridge, 2007). In the mid 1990s, phasic

dopamine activations were recognized to correspond closely with the behavior of the Temporal Difference prediction error (TD error) postulated by the TD algorithm of computational reinforcement learning (Barto, 1995; Houk et al., 1995; Schultz et al., 1997; Schultz, 1998). This has led to the reward-prediction-error hypothesis of the phasic activity of dopamine neurons, which has received a large amount of empirical support and represents one of the most fruitful integrations between computational and empirical research (Ungless, 2004; Wise, 2004; Schultz, 2007; Graybiel, 2008; Glimcher, 2011).

Notwithstanding its success, an important problem faced by the reward-prediction-error hypothesis is that phasic dopamine neuron activity is not triggered only by rewards and reward predictors, but by different kinds of salient stimuli (Horvitz, 2000), such as sudden visual or auditory stimuli that have never been associated with rewards (Steinfels et al., 1983; Ljungberg et al., 1992; Horvitz et al., 1997). Because these responses tend to disappear with repeated stimulation, they have been called “novelty” responses (Schultz, 1998). An interesting explanation of these responses has been proposed by Kakade and Dayan (2002b), who relate them to the problem of exploration: according to these authors, these dopamine activations represent “novelty bonuses” that are generated when an animal perceives novel states and that serve the function of increasing the animal’s tendency to explore the environment, thus augmenting the probability that the animal finds rewards. The novelty bonus idea has recently attracted much attention, and it is fostering a number of neuroimaging studies where the activation of the dopaminergic system is studied while subjects are exposed to novel stimuli (e.g., Bunzeck and Duzel, 2006; Wittmann et al., 2008; Krebs et al., 2009).

The problem here is that the so-called novelty responses of dopamine neurons found in animals through electrophysiological studies do not seem to be related to novelty, but rather to surprise. In fact, the stimuli that have been used in those electrophysiological experiments are simple light flashes or sudden sounds, and the dopaminergic responses to lights and tones typically persist after many presentations so that talking about novelty of the stimuli does not seem appropriate (Steinfels et al., 1983; Horvitz et al., 1997; Ungless, 2004). Hence, it is more reasonable to assume that it is the *unexpectedness* of the event, e.g., the sudden appearance of a light or sound, that is responsible for dopamine activation.

Further indirect evidence that the activity of dopaminergic neurons triggered by lights and tones is due to surprise rather than novelty comes from behavioral studies of sensory reinforcement. Sensory reinforcement is the very well-investigated phenomenon that many kinds of sensory events (of which the most frequently studied are again lights and tones) are able to drive the acquisition of instrumental responses. For example, if pressing a bar results in the switching on of a light, an animal will start to press the bar, much as if the bar-press were to lead to a reward such as food (e.g., Kish, 1955; Williams and Lowe, 1972; Glow and Winefield, 1978; Reed et al., 1996). Because we know that dopamine is both necessary and sufficient for appetitive instrumental conditioning (Robinson et al., 2006; Zweifel et al., 2009), it is probably safe to assume that it is phasic dopamine that mediates operant conditioning in sensory reinforcement, just as

we assume that it is dopamine that drives standard instrumental conditioning reinforced by food.

Further support that surprise and not novelty supports sensory reinforcement comes from the evidence that light offsets are more-or-less as good reinforcers as light onsets (Glow, 1970; Russell and Glow, 1974). But in the case of light offset, where is the “novel” stimulus that acts as a reinforcer (by supposedly triggering dopamine)? In this case it is even more clear that it is the unexpectedness of the event (surprise), not the novelty of the stimulus (which is absent), that is at play.

We have argued that it is surprise and not novelty that triggers phasic activity of dopamine neurons in animal electrophysiological studies involving lights and tones. But why should this mere misuse of terminology be worth noting? We think there are at least two important reasons to be aware of this misleading labeling. The first reason has to do with the mechanisms underlying phasic activation of dopamine neurons. If one wants to understand how dopamine neuron activity is triggered, it is probably a good idea not to confuse novelty activations due to novel images with surprise activations due to unexpected events. In fact, not surprisingly in human experiments with novel images, it is the hippocampus that seems to be involved (e.g., Lisman and Grace, 2005), whereas light flashes trigger dopamine activity via the superior colliculus, which directly projects to the dopaminergic neurons (Dommett et al., 2005). Furthermore, if it is the unexpectedness of lights or tones that trigger dopamine neuron activity, then the question is raised about the neural circuits providing the predictions that inhibit surprise activations after repeated stimulation. This is a very important question that, to the best of our knowledge, has not yet been addressed. We conjecture that a key reason for this neglect is that these dopamine responses have been regarded as novelty responses, and therefore that they do not involve predictions.

The second reason the novelty/surprise distinction is important with respect to phasic activity of dopamine neurons has to do with the function that these activations play in animal behavior. While it is reasonable to assume that the dopaminergic responses to novel stimuli found in animals are actually “novelty bonuses” that facilitate exploration (Kakade and Dayan, 2002b), it is less reasonable to assume that the same function is ascribed to dopamine activations triggered by unexpected (surprising) events. In fact, it seems more likely that the function of dopamine surprise activations is to encourage the animal to engage in activity to discover which aspects of its own activity may trigger surprising events so that the animal may add new actions to its repertoire (Redgrave et al., 1999; Redgrave and Gurney, 2006; Mirolli et al., 2013).

Finally, to reiterate a point made in Section 2, the TD algorithm, which underlies the reward-prediction-error hypothesis of phasic dopamine neuron activity, is not restricted to predicting reward: the role of reward can be replaced by other stimulus features. The reward-prediction-error hypothesis essentially says that the TD error signals the surprising receipt of reward. But the same machinery equally can signal the surprising receipt of any stimulus. As in the Rescorla-Wagner model, the essence of TD learning is surprise. This adds further support to our suggestion that it would be better to think of the phasic activity

of dopamine neurons as responses to surprise rather than to novelty.

7. CONCLUSION

Novelty and surprise play significant roles in animal behavior and in attempts to understand the neural mechanisms underlying it. Surprise and novelty underlie core intrinsic motivations that allow organisms (and promise to allow robots) to acquire useful knowledge and skills in the absence of explicit instruction and externally supplied rewards and penalties. They also play important roles in technology, where detecting observations that are novel or surprising is central to many applications, such as medical diagnosis, text processing, surveillance, and security. The words novelty and surprise are often used interchangeably despite the fact that according to our normal understanding novelty and surprise refer to very different phenomena. Without claiming to do justice to all that has been written about novelty and surprise, we described a sample of past attempts to define these concepts, and we related these definitions to our common sense notions. We pointed out key factors distinguishing surprise from novelty, and we argued that some of the definitions in common use are misleading, as are some of the labels and interpretations applied to results of experiments by psychologists and neuroscientists. But clarifying, indeed in some cases correcting, word usage has not been our goal: opportunities for improved understanding of behavior and its neural basis are likely being missed by failing to distinguish between novelty and surprise.

ACKNOWLEDGMENTS

The authors thank Barak Pearlmuter for pointing out Huron's PDQ Bach example, Ashvin Shah for helping us track down many references, John Moore for his sage input, and anonymous reviewers for their very helpful suggestions. This research was funded by the European Community 7th Framework Programme (FP7/2007-2013), "Challenge 2—Cognitive Systems, Interaction, Robotics," grant agreement No. ICT-IP-231722, project "IM-CLeVeR—Intrinsically Motivated Cumulative Learning Versatile Robots."

REFERENCES

- Asaad, W. F., Rainer, G., and Miller, E. K. (1998). Neural activity in the primate prefrontal cortex during associative learning. *Neuron* 21, 1399–1407. doi: 10.1016/S0896-6273(00)80658-3
- Baldassarre, G., and Mirolli, M. (eds.). (2013). *Intrinsically Motivated Learning in Natural and Artificial Systems*. Berlin: Springer-Verlag. doi: 10.1007/978-3-642-32375-1
- Barto, A. G. (1990). "Connectionist learning for control: an overview," in *Neural Networks for Control*, eds T. Miller, R. S. Sutton, and P. J. Werbos (Cambridge, MA: MIT Press), 5–58.
- Barto, A. G. (1995). "Adaptive critics and the basal ganglia," in *Models of Information Processing in the Basal Ganglia*, eds J. C. Houk, J. L. Davis, and D. G. Beiser (Cambridge, MA: MIT Press), 215–232.
- Berlyne, D. E. (1960). *Conflict, Arousal, and Curiosity*. New York, NY: McGraw-Hill. doi: 10.1037/11164-000
- Berridge, K. (2007). The debate over dopamine's role in reward: the case for incentive salience. *Psychopharmacology* 191, 391–431. doi: 10.1007/s00213-006-0578-x
- Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. New York, NY: Springer.
- Brown, M. W., and Aggleton, J. P. (2001). Recognition memory: what are the roles of the perirhinal cortex and hippocampus? *Nat. Rev. Neurosci.* 2, 51–61. doi: 10.1038/35049064
- Bunzeck, N., and Duzel, E. (2006). Absolute coding of stimulus novelty in the human substantia nigra/vta. *Neuron* 51, 369–379. doi: 10.1016/j.neuron.2006.06.021
- Courville, A. C., Daw, N. D., Gordon, G. J., and Touretzky, D. S. (2004). "Model uncertainty in classical conditioning," in *Advances in Neural Information Processing Systems 16*, eds S. Thrun, L. Saul, and B. Schölkopf (Cambridge, MA: MIT Press), 977–984.
- Courville, A. C., Daw, N. D., and Touretzky, D. S. (2006). Bayesian theories of conditioning in a changing world. *Trends Cogn. Sci.* 10, 294–300. doi: 10.1016/j.tics.2006.05.004
- Dayan, P., Kakade, S., and Montague, P. R. (2000). Learning and selective attention. *Nat. Neurosci. Suppl.* 3, 1218–1223. doi: 10.1038/81504
- Dayan, P., and Long, T. (1998). "Statistical models of learning," in *Advances in Neural Information Processing Systems 10: Proceedings of the 1997 Conference*, eds M. I. Jordan, M. J. Kearns, and S. A. Solla (Cambridge, MA: MIT Press), 117–123.
- Deci, E., and Ryan, R. (1985). *Intrinsic Motivation and Self-Determination in Human Behavior*. New York, NY: Plenum Press. doi: 10.1007/978-1-4899-2271-7
- Dommett, E., Coizet, V., Blaha, C. D., Martindale, J., Lefebvre, V., Walton, N., et al. (2005). How visual stimuli activate dopaminergic neurons at short latency. *Science* 307, 1476–1479. doi: 10.1126/science.1107026
- Duda, R. O., and Hart, P. E. (1973). *Pattern Classification and Scene Analysis*. New York, NY: Wiley.
- Duncan, K., Ketz, N., Inati, S. J., and Davachi, L. (2012). Evidence for area cal as a match/mismatch detector: a high-resolution fmri study of the human hippocampus. *Hippocampus* 22, 389–398. doi: 10.1002/hipo.20933
- Duzel, E., Habib, R., Rotte, M., Guderian, S., Tulving, E., and Heinze, H. (2003). Human hippocampal and parahippocampal activity during visual associative recognition memory for spatial and nonspatial stimulus configurations. *J. Neurosci.* 23, 9439–9444. Available online at: <http://www.jneurosci.org/content/23/28/9439.long>
- Ekman, P., and Davidson, R. J. (eds.). (1994). *The Nature of Emotion: Fundamental Questions*. Oxford: Oxford University Press.
- Engle, Y., Mannor, S., and Meir, R. (2003). "Bayes meets Bellman: the Gaussian process approach to temporal difference learning," in *Proceedings of the twentieth International Conference on Machine Learning (ICML-2003)* (Washington, DC), 154–161.
- Friedman, D., Cycowicz, Y. M., and Gaeta, H. (2001). The novelty p3: an event-related brain potential (erp) sign of the brain's evaluation of novelty. *Neurosci. Biobehav. Rev.* 25, 355–373. doi: 10.1016/S0149-7634(01)00019-7
- Friston, K. (2009). The free-energy principle: a rough guide to the brain? *Trends Cogn. Sci.* 13, 293–301. doi: 10.1016/j.tics.2009.04.005
- Friston, K. (2010). The free-energy principle: a unified brain theory? *Nat. Rev. Neurosci.* 11, 127–138. doi: 10.1038/nrn2787
- Friston, K., and Kiebel, S. (2009). Predictive coding under the free-energy principle. *Philos. Trans. R. Soc. B* 364, 1211–1221. doi: 10.1098/rstb.2008.0300
- Friston, K. J., Kilner, J., and Harrison, L. (2006). A free-energy principle for the brain. *J. Physiol. Paris* 100, 70–87. doi: 10.1016/j.jphysparis.2006.10.001
- Gershman, S. J., and Blei, D. M. (2012). A tutorial on bayesian nonparametric models. *J. Math. Psychol.* 56, 1–12. doi: 10.1016/j.jmp.2011.08.004
- Glimcher, P. (2011). Understanding dopamine and reinforcement learning: the dopamine reward prediction error hypothesis. *Proc. Natl. Acad. Sci. U.S.A.* 108(Suppl. 3), 15647–15654. doi: 10.1073/pnas.1014269108
- Glow, P. (1970). Some acquisition and performance characteristics of response contingent sensory reinforcement in the rat. *Aust. J. Psychol.* 22, 145–154. doi: 10.1080/00049537008254568
- Glow, P., and Winefield, A. (1978). Response-contingent sensory change in a causally structured environment. *Learn. Behav.* 6, 1–18. doi: 10.3758/BF03211996
- Graybiel, A. (2008). Habits, rituals, and the evaluative brain. *Annu. Rev. Neurosci.* 31, 359–387. doi: 10.1146/annurev.neuro.29.051605.112851
- Grossberg, S. (1982). Processing of expected and unexpected events during conditioning and attention: a psychophysiological theory. *Psychol. Rev.* 89, 529–572. doi: 10.1037/0033-295X.89.5.529

- Hasselmo, M. E., and Schnell, E. (1994). Laminar selectivity of the cholinergic suppression of synaptic transmission in rat hippocampal region cal: computational modeling and brain slice physiology. *J. Neurosci.* 14, 3898–3914.
- Henson, R. N. A., and Rugg, M. D. (2003). Neural response suppression, haemodynamic repetition effects, and behavioural priming. *Neuropsychologia* 41, 263–270. doi: 10.1016/S0028-3932(02)00159-8
- Horvitz, J. C. (2000). Mesolimbocortical and nigrostriatal dopamine responses to salient non-reward events. *Neuroscience* 96, 651–656. doi: 10.1016/S0306-4522(00)00019-1
- Horvitz, J. C., Stewart, T., and Jacobs, B. L. (1997). Burst activity of ventral tegmental dopamine neurons is elicited by sensory stimuli in the awake cat. *Brain Res.* 759, 251–258. doi: 10.1016/S0006-8993(97)00265-5
- Houk, J. C., Adams, J. L., and Barto, A. G. (1995). “A model of how the basal ganglia generates and uses neural signals that predict reinforcement,” in *Models of Information Processing in the Basal Ganglia*, eds J. C. Houk, J. L. Davis, and D. G. Beiser (Cambridge, MA: MIT Press), 249–270.
- Huron, D. (2004). “Music-engendered laughter: an analysis of humor devices in PDQ Bach,” in *Proceedings of the 8th International Conference on Music Perception and Cognition*, eds S. D. Lipscomb, R. Ashley, R. O. Gjerdingen, and P. Webster (Adelaide, SA: Causal Productions), 700–704.
- Itti, L., and Baldi, P. F. (2005). “A principled approach to detecting surprising events in video,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (San Diego, CA), 631–637.
- Itti, L., and Baldi, P. F. (2006). “Bayesian surprise attracts human attention,” in *Advances in Neural Information Processing Systems 18 (NIPS*2005)*, eds Y. Weiss, B. Schölkopf, and J. Platt (Cambridge, MA: MIT Press), 547–554.
- Itti, L., and Baldi, P. F. (2009). Bayesian surprise attracts human attention. *Vis. Res.* 49, 1295–1306. doi: 10.1016/j.visres.2008.09.007
- Kakade, S., and Dayan, P. (2002a). Acquisition and extinction in autoshaping. *Psychol. Rev.* 109, 533–544. doi: 10.1037/0033-295X.109.3.533
- Kakade, S., and Dayan, P. (2002b). Dopamine: generalization and bonuses. *Neural Netw.* 15, 549–559. doi: 10.1016/S0893-6080(02)00048-5
- Kamin, L. J. (1969). “Predictability, surprise, attention, and conditioning,” in *Punishment and Aversive Behavior*, eds B. A. Campbell and R. M. Church (New York, NY: Appleton-Century-Crofts), 279–296.
- Kish, G. B. (1955). Learning when the onset of illumination is used as reinforcing stimulus. *J. Comp. Physiol. Psychol.* 48, 261–264. doi: 10.1037/h0040782
- Kohler, S., Danckert, S., Gati, J., and Menon, R. (2005). Novelty responses to relational and non-relational information in the hippocampus and the parahippocampal region: a comparison based on event-related fmri. *Hippocampus* 15, 763–774. doi: 10.1002/hipo.20098
- Kohonen, T. (1977). *Associative Memory: A System Theoretic Approach*. Berlin: Springer-Verlag.
- Kohonen, T. (1980). *Content-Addressable Memories*. Berlin: Springer-Verlag. doi: 10.1007/978-3-642-96552-4
- Kohonen, T. (1984). *Self-Organization and Associative Memory*. Berlin: Springer-Verlag.
- Krebs, R. M., Schott, B. H., Schutze, H., and Duzel, E. (2009). The novelty exploration bonus and its attentional modulation. *Neuropsychologia* 47, 2272–2281. doi: 10.1016/j.neuropsychologia.2009.01.015
- Kumaran, D., and Maguire, E. A. (2007). Which computational mechanisms operate in the hippocampus during novelty detection? *Hippocampus* 17, 735–748. doi: 10.1002/hipo.20326
- Lepora, N. F., Porrill, J., Yeo, C. H., and Dean, P. (2010). Sensory prediction or motor control? Application of Marr-Albus type models of cerebellar function to classical conditioning. *Front. Comput. Neurosci.* 4:140. doi: 10.3389/fncom.2010.00140
- Li, L., Miller, E. K., and Desimone, R. (1993). The representation of stimulus familiarity in anterior inferior temporal cortex. *J. Neurophysiol.* 69, 1918–1929.
- Lisman, J. E., and Grace, A. A. (2005). The hippocampal-vta loop: controlling the entry of information into long-term memory. *Neuron* 46, 703–713. doi: 10.1016/j.neuron.2005.05.002
- Ljungberg, T., Apicella, P., and Schultz, W. (1992). Responses of monkey dopamine neurons during learning of behavioral reactions. *J. Neurophysiol.* 67, 145–163.
- Mannella, F., Zappacosta, S., Mirolli, M., and Baldassarre, G. (2010). “A computational model of the amygdala nuclei’s role in second order conditioning,” in *From Animals to Animats 10: Proceedings of the Tenth International Conference on the Simulation of Adaptive behavior (SAB2008), Lecture Notes in Artificial Intelligence 5040*, eds M. Asada, J. C. Hallam, J.-A. Meyer, and J. Tani (Berlin: Springer-Verlag).
- Markou, M., and Singh, S. (2003). Novelty detection: a review - part 1: statistical approaches. *Signal Process.* 83, 2481–2497. doi: 10.1016/j.sigpro.2003.07.018
- Marsland, S. (2003). Novelty detection in learning systems. *Neural Comput. Surv.* 3, 157–195. Available online at: <http://seat.massey.ac.nz/personal/s.r.marsland/PUBS/NCS.pdf>
- Mathys, C., Daunizeau, J., Friston, K. J., and Stephan, K. E. (2011). A Bayesian foundation for individual learning under uncertainty. *Front. Hum. Neurosci.* 5:39. doi: 10.3389/fnhum.2011.00039
- Mirolli, M., Santucci, V., and Baldassarre, G. (2013). Phasic dopamine as a prediction error signal of intrinsic and extrinsic reinforcements: a computational model. *Neural Netw.* 39, 40–51. doi: 10.1016/j.neunet.2012.12.012
- Monsalve, I. F., Frank, S. L., and Vigliocco, G. (2012). “Lexical surprisal as a general predictor of reading time,” in *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics* (Stroudsburg, PA: Association for Computational Linguistics), 398–408.
- Moore, J. W., and Schmajuk, N. A. (2008). Kamin blocking. *Scholarpedia* 3, 3542. doi: 10.4249/scholarpedia.3542
- Nehmzow, U., Gatsoulis, Y., Kerr, E., Condell, J., Siddique, N., and McGinnity, T. (2013). “Novelty detection as an intrinsic motivation for cumulative learning robots,” in *Intrinsically Motivated Learning in Natural and Artificial Systems*, eds G. Baldassarre and M. Mirolli (Berlin: Springer-Verlag), 185–207.
- O’Keefe, J., and Nadel, L. (1978). *The Hippocampus as a Cognitive Map*. Oxford: Oxford University Press.
- Palm, G. (2012). *Novelty, Information and Surprise*. Berlin: Springer-Verlag. doi: 10.1007/978-3-642-29075-6
- Pearce, J. M., and Hall, G. (1980). A model for Pavlovian learning: variation in the effectiveness of conditioning but not unconditioned stimuli. *Psychol. Rev.* 87, 532–552. doi: 10.1037/0033-295X.87.6.532
- Ranganath, C., and Paller, K. A. (1999). Frontal brain activity during episodic and semantic retrieval: insights from event-related potentials. *J. Cogn. Neurosci.* 11, 598–609. doi: 10.1162/08989299563661
- Ranganath, C., and Rainer, G. (2003). Neural mechanisms for detecting and remembering novel events. *Nat. Rev. Neurosci.* 4, 193–202. doi: 10.1038/nrn1052
- Rao, R. P., and Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.* 2, 79–87. doi: 10.1038/4580
- Redgrave, P., and Gurney, K. (2006). The short-latency dopamine signal: a role in discovering novel actions? *Nat. Rev. Neurosci.* 7, 967–975. doi: 10.1038/nrn2022
- Redgrave, P., Prescott, T. J., and Gurney, K. (1999). Is the short-latency dopamine response too short to signal reward error? *Trends Neurosci.* 22, 146–151. doi: 10.1016/S0166-2236(98)01373-3
- Reed, P., Mitchell, C., and Nokes, T. (1996). Intrinsic reinforcing properties of putatively neutral stimuli in an instrumental two-lever discrimination task. *Anim. Learn. Behav.* 24, 38–45. doi: 10.3758/BF03198952
- Rescorla, R. A. (1971). Variations in the effectiveness of reinforcement and nonreinforcement following prior inhibitory conditioning. *Learn. Motiv.* 2, 113–123. doi: 10.1016/0023-9690(71)90002-6
- Rescorla, R. A., and Wagner, A. R. (1972). “A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement,” in *Classical Conditioning II*, eds A. H. Black and W. F. Prokasy (New York, NY: Appleton-Century-Crofts), 64–99.
- Ringo, J. L. (1996). Stimulus specific adaptation in inferior temporal and medial temporal cortex of the monkey. *Behav. Brain Res.* 76, 191–197. doi: 10.1016/0166-4328(95)00197-2
- Roark, B. (2011). *Expected Surprisal and Entropy*. Technical Report CSLU-11-004, Center for Spoken Language Processing, Oregon Health and Science University, (Portland, OR).
- Robinson, S., Sotak, B., During, M., and Palmiter, R. (2006). Local dopamine production in the dorsal striatum restores goal-directed behavior in dopamine-deficient mice. *Behav. Neurosci.* 120, 196–200. doi: 10.1037/0735-7044.120.1.000
- Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). “Learning internal representations by error propagation,” in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1: Foundations*, eds D. E. Rumelhart and J. L. McClelland (Cambridge, MA: Bradford Books/MIT Press), 318–362.

- Russell, A., and Glow, P. (1974). Some effects of short-term immediate prior exposure to light change on responding for light change. *Learn. Behav.* 2, 262–266. doi: 10.3758/BF03199191
- Schmajuk, N. A. (2008). Computational models of classical conditioning. *Scholarpedia* 3, 1664. doi: 10.4249/scholarpedia.1664
- Schmidhuber, J., Storck, J., and Hochreiter, S. (1994). *Reinforcement driven information acquisition in nondeterministic environments*. Munich: Technical report, Fakultät für Informatik, Technische Universität München.
- Schultz, W. (1998). Predictive reward signal of dopamine neurons. *J. Neurophysiol.* 80, 1–27.
- Schultz, W. (2007). Multiple dopamine functions at different time scales. *Annu. Rev. Neurosci.* 30, 259–288. doi: 10.1146/annurev.neuro.28.061604.135722
- Schultz, W., Dayan, P., and Montague, P. R. (1997). A neural substrate of prediction and reward. *Science* 275, 1593–1598. doi: 10.1126/science.275.5306.1593
- Sobotka, S., and Ringo, J. L. (1994). Stimulus specific adaptation in excited but not inhibited cells in inferotemporal cortex of macaque. *Brain Res.* 646, 95–99. doi: 10.1016/0006-8993(94)90061-2
- Soltani, M., and Knight, R. T. (2000). Neural origins of the p300. *Crit. Rev. Neurobiol.* 14, 199–224. doi: 10.1615/CritRevNeurobiol.v14.i3-4.20
- Steinfels, G. F., Heym, J., Strecker, R. E., and Jacobs, B. L. (1983). Response of dopaminergic neurons in cat to auditory stimuli presented across the sleep-waking cycle. *Brain Res.* 277, 150–154. doi: 10.1016/0006-8993(83)90917-4
- Storck, J., Hochreiter, S., and Schmidhuber, J. (1995). “Reinforcement-driven information acquisition in non-deterministic environments,” in *Proceedings of ICANN’95* (Paris), Vol. 2, 159–164.
- Sutton, R. S. (1988). Learning to predict by the method of temporal differences. *Mach. Learn.* 3, 9–44. doi: 10.1007/BF00115009
- Sutton, R. S., and Barto, A. G. (1990). “Time-derivative models of Pavlovian reinforcement,” in *Learning and Computational Neuroscience: Foundations of Adaptive Networks*, eds M. Gabriel and J. Moore (Cambridge, MA: MIT Press), 497–537.
- Sutton, R. S., and Tanner, B. (2004). “Temporal-difference networks,” in *Advances in Neural Information Processing Systems 17, [Neural Information Processing Systems, NIPS 2004]*, (Vancouver, BC), 1377–1384.
- Tribus, M. (1961). *Thermodynamics and Thermostatics: An Introduction to Energy, Information and States of Matter, with Engineering Applications*. New York, NY: D. Van Nostrand Company Inc.
- Ungless, M. (2004). Dopamine: the salient issue. *Trends Neurosci.* 27, 702–706. doi: 10.1016/j.tins.2004.10.001
- Wan, H., Aggleton, J. P., and Brown, M. W. (1999). Different contributions of the hippocampus and perirhinal cortex to recognition memory. *J. Neurosci.* 19, 1142–1148.
- Welch, G., and Bishop, G. (1995). *An introduction to the kalman filter*. Technical report, Department of Computer Science, University of North Carolina at Chapel Hill, (Chapel Hill, NC).
- Widrow, B., and Hoff, M. E. (1960). “Adaptive switching circuits,” in *1960 WESCON Convention Record Part IV*, (NY: Institute of Radio Engineers) 96–104, Reprinted in Anderson, J. A., and Rosenfeld, E. (1988). *Neurocomputing: Foundations of Research*, (Cambridge, MA: MIT Press), 126–134.
- Williams, D., and Lowe, G. (1972). Response contingent illumination change as a reinforcer in the rat. *Anim. Behav.* 20, 259–262. doi: 10.1016/S0003-3472(72)80045-9
- Wise, R. (2004). Dopamine, learning and motivation. *Nat. Rev. Neurosci.* 5, 483–494. doi: 10.1038/nrn1406
- Wittmann, B. C., Daw, N. D., Seymour, B., and Dolan, R. J. (2008). Striatal activity underlies novelty-based choice in humans. *Neuron* 58, 967–973. doi: 10.1016/j.neuron.2008.04.027
- Zweifel, L. S., Parker, J. G., Lobb, C. J., Rainwater, A., Wall, V. Z., Fadok, J. P., et al. (2009). Disruption of nmdar-dependent burst firing by dopamine neurons provides selective assessment of phasic dopamine-dependent behavior. *Proc. Natl. Acad. Sci. U.S.A.* 106, 7281–7288. doi: 10.1073/pnas.0813415106

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 05 July 2013; paper pending published: 23 September 2013; accepted: 15 November 2013; published online: 11 December 2013.

*Citation: Barto A, Mirolli M and Baldassarre G (2013) Novelty or Surprise? *Front. Psychol.* 4:907. doi: 10.3389/fpsyg.2013.00907*

This article was submitted to Cognitive Science, a section of the journal Frontiers in Psychology.

Copyright © 2013 Barto, Mirolli and Baldassarre. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Learning autonomy in two or three steps: linking open-ended development, authority, and agency to motivation

Tjeerd C. Andringa^{1*}, Kirsten A. van den Bosch² and Carla Vlaskamp²

¹ Artificial Intelligence and Cognitive Engineering (ALICE), University of Groningen, Groningen, Netherlands

² Special Needs Education and Youth Care, University of Groningen, Groningen, Netherlands

Edited by:

Gianluca Baldassarre, Italian National Research Council, Italy

Reviewed by:

Matthew Schlesinger, Southern Illinois University, USA

Geoffrey Chern-Yee Tan, National Healthcare Group, Singapore

Kathryn E. Merrick, University of New South Wales, Australia

***Correspondence:**

Tjeerd C. Andringa, Artificial Intelligence and Cognitive Engineering, University of Groningen, PO Box 407, 9700 AK Groningen, Netherlands
e-mail: tjeerd@ai.rug.nl

In this paper we connect open-ended development, authority, agency, and motivation through (1) an analysis of the demands of existing in a complex world and (2) environmental appraisal in terms of affordance content and the complexity to select appropriate behavior. We do this by identifying a coherent core from a wide range of contributing fields. Open-ended development is a structured three-step process in which the agent first learns to master the body and then aims to make the mind into a reliable tool. Preconditioned on success in step two, step three aims to effectively co-create an optimal living environment. We argue that these steps correspond to right-left-right hemispheric dominance, where the left hemisphere specializes in control and the right hemisphere in exploration. Control (e.g., problem solving) requires a closed and stable world that must be maintained by external authorities or, in step three, by the right hemisphere acting as internal authority. The three-step progression therefore corresponds to increasing autonomy and agency. Depending on how we appraise the environment, we formulate four qualitatively different motivational states: submission, control, exploration, and consolidation. Each of these four motivational states has associated reward signals of which the last three—successful control, discovery of novelty, and establishing new relations—form an open-ended development loop that, the more it is executed, helps the agent to become progressively more agentic and more able to co-create a pleasant-to-live-in world. We conclude that for autonomy to arise, the agent must exist in a (broad) transition region between order and disorder in which both danger and opportunity (and with that open-ended development and motivation) are defined. We conclude that a research agenda for artificial cognitive system research should include open-ended development through intrinsic motivations and ascribing more prominence to right hemispheric strengths.

Keywords: motivation, agency, autonomy, open-ended development, co-creation, authority, complexity, lateralization

INTRODUCTION

In this theoretical paper we aim to unify a number of complementary and highly consistent results from a wide range of scientific domains that all pertain to “learning to cope autonomously with the challenges of an open environment.” We will frame these results in terms of agency and autonomy development. In the final section we will formulate what we call the “open-ended development loop” (Figure 5) as a main and productive synthesis for artificial cognitive system research and behavioral sciences in general.

In our efforts we benefitted from results and insights from life-span research, personality development, emotion theory, psychoanalysis, motivation research, brain lateralization, political psychology, soundscape research, complexity theory, and even early Chinese philosophy. In addition, although in this paper less prominent, we benefited from moral psychology, epistemological development, and education research. While this may

seem an unnecessary wide range of scientific domains to address the call-topic of “open-ended development driven by intrinsic motivations” we argue that both the concepts of “open-ended development” and “motivation” are not just cognitive functions, but cognitive foundations: without motivation there would be no activity and no agency.

As cognitive foundations, “motivation” and “open-ended development” shape and constrain many facets of cognition. As such, insights from all specialisms of the cognitive sciences in the broadest sense, and in particular those domains directly related to open environments, may contribute with novel perspectives on foundational principles. We will outline that open-ended development and motivation are intimately related with concepts such as agency, mood, behavior, and action selection, brain lateralization, appraisal, safety, and complexity. In addition we will introduce the terms “authority” (defined as the capacity to create, maintain, and influence living environments), and “co-creation”

(defined as the ability to work with the inherent dynamics of the world instead of suppressing and controlling it) as fundamental concepts for understanding agency and cognition.

Since we derive from many domains of science we focus more on the relations between relevant concepts and the progression of argument than on experimental or implementation details. In many cases we will slightly generalize domain specific terms, concepts, and results to make them more consistent with each other. Our inductive approach to science was only possible because of the many deep and precisely formulated insights by researchers from very different traditions, which strengthens our belief that the true value of scientific insights can only be estimated outside of the domain where it was developed. We present many of these insights as direct quotes so that the quality of the formulation can also be appreciated in a quite different context than the original publication.

Our paper has, apart from the introduction, 4 main sections. In section Open-Ended Development we address a wide range of results and insights consistent with the title of our paper, suggesting that open-ended development occurs in two or three steps, with the third step being pre-conditioned on the success of the second. In step one the agent's focus is on making the body into a reliable instrument. The second step involves making the mind into a reliable and effective tool. Only success in this step allows a third phase in which the agent learns to shape—co-create—the conditions for its continued existence and in doing so it becomes independent of external authority and truly autonomous. We visualize this two or three step approach as a spiral development in which matching development phases stemming from diverse fields of research have been indicated. This spiral epitomizes open-ended development.

In the next section we address two attitudes toward a complex world. One associated with exploration and the other with control. We couple these attitudes to two modes of being and understanding of the world that comply very well with the different strengths of the left (control) and right (exploration) hemisphere. Here we conclude that step one and three rely on right hemispheric dominance and step two on left hemispheric dominance. We couple this conclusion to a need of external authority associated with a dominant left hemisphere.

In section Motivations we address motivations by first focusing on some of our own results that couple four qualitatively different appraisals of the (sonic) environment to motivational states in the context of core affect. We argue that each of the four quadrants of core affect constrains mind-states in a distinct way and that motivation can be treated as attitudes toward particularly appraised worlds. We end this section with a table describing these four quadrants in terms of motivation and other properties derived from different scientific domains.

In section Open-Ended Development Driven by Intrinsic Motivation we address the call topic “open-end development driven by intrinsic motivations” by outlining the conditions required for open-ended development, which, we argue, rely essentially on the agent learning to shape its own environment. We argue that the results of motivation research, interpreted in the context of the four quadrants, describe what we call the “open-ended development loop.” We note a number of observations and

constraints to be satisfied for open-end development to occur that might be used to formulate a research agenda for artificial cognitive systems research. We end with the observation that particular Western—left hemispheric—biases have limited our understanding of cognitive systems and we suggest a way to address these limitations.

OPEN-ENDED DEVELOPMENT

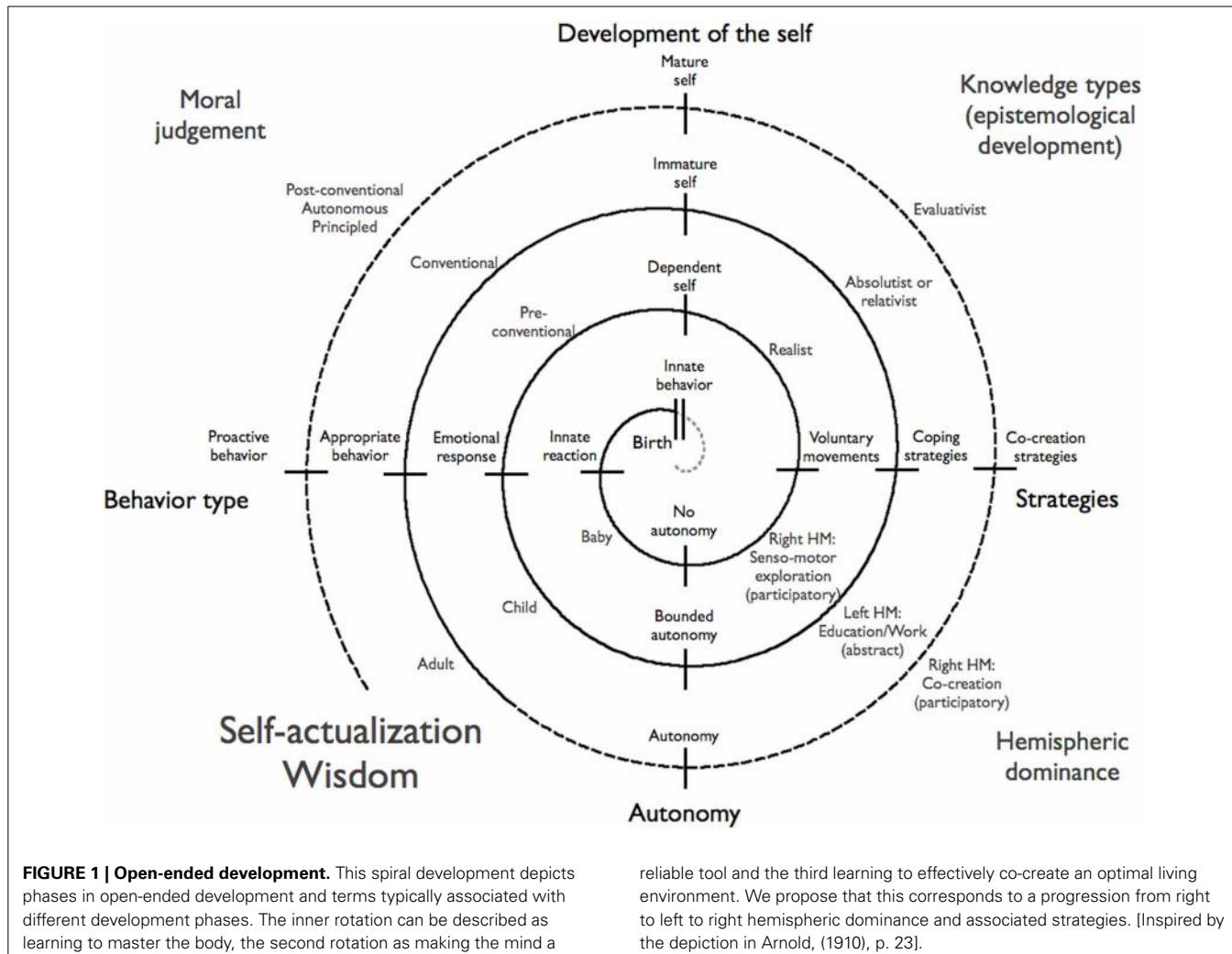
Open-ended development is not undirected, quite the contrary. The research outlined below shows that open-ended development refers to the capacity to ever-extend and fine-tune one's capacity to deal with life's challenges and to co-create one's environment. Put differently: open-ended development is a development that allows agents to gradually master more and more of the complexity of the world and to become more and more self-deciding, agentic, and autonomous. **Figure 1** visualizes open-ended development and it summarizes many of the results that we address in this section in terms of reported stadia of open-ended development. While this section addresses the many properties of open-ended development, its main drivers—the demands of an open world, (intrinsic) motivation, and the open-ended development loop—will be addressed in later sections.

The spiral development outwards makes about three turns that reflect, very roughly, three developmental phases. The first phase is physical growth and learning to control the body. In the second phase one aims to make the mind into a reliable instrument. The third phase, preconditioned on the success of phase 2, concerns learning to co-construct a world in which the inherent dynamics of the world are stabilized and made reliable and broadly beneficial. This leads to ever more extended (both in place and in time) environments in which one can self-maintain the condition for adequate functioning, leading to increasing diversity and individual authority. This characterizes the outer (pre-conditioned) loop of the spiral development in **Figure 1**.

The figure has a number of functional components. The spiral is divided into a number of sectors that reflect aspects of open-ended development without being necessarily in the strict circular progression the spiral form suggests. The end-state of the spiral is referred to as self-actualization or wisdom. The solid-line part of the spiral reflects development up the level of the authoritarian personality, while the dashed part reflects an—in Western cultures non-standard—additional development toward the libertarian personality type. The main axes reflect behavior types and strategies horizontally and self-development and action readiness vertically. The diagonal axes reflect distinct development stages from diverse scientific fields: moral development, education research and epistemological development, and brain lateralization research. In the next subsections we'll provide supportive evidence for each axis and gradually develop the key terminology of this paper.

END-STATE: SELF-ACTUALIZATION AND WISDOM

Open-ended development in humans is a highly structured process that has been well studied in a variety of different domains that each shed more light on the phases in the development process. The development process begins obviously at conception and develops after birth in a number of stages toward



what Maslow (1943, 1962) calls self-actualization. According to Maslow, self-actualization accounts for the highest possible forms of psychological health and self-development. As such it is a candidate for fully developed open-ended learning. Among the main characteristic properties of a self-actualized individual are (1) realistic perceptions of themselves, others, and the world around them, (2) a strong motivation, through a sense of personal responsibility and ethics, to help others and to find solutions to problems in the external world, and (3) a well-developed personal autonomy, which is for example visible as an utter disregard of conformity if the situation demands this and an appreciation for private time to self-develop one's potential further.

Compared to not (yet) self-actualized individuals they

1. Have learned the skills to prevent or overcome one's own psychological problems that allow them to be rarely motivated by unfulfilled needs,
2. Have developed a deep and pervasive understanding of reality that they keep extending through life and that is apparent from a well-developed creative capacity to produce intended results with minimal adverse side-effects, and

3. Feel a moral obligation to contribute to an improved world.

These properties reflect deep realities concerning the nature of agentic life. Interestingly the term self-actualization arose from Maslow's work on motivation (Maslow, 1943), but he refined and defined the term self-actualization later on the basis of case-studies of individuals of whom he thought that they represented examples of self-actualization (Maslow, 1962). This intuition-driven (dangerously circular) process is vindicated by results later in this paper that dovetail with Maslow's conclusions while being based on entirely different evidence.

Another way to approach open-ended development comes from gerontology and especially the role of lifelong learning and continued education for older people which allows them to stay involved in a rapidly changing world (Ardelt, 2000). This led to a distinction between intellectual knowledge and wisdom-related knowledge, of which the wisdom related knowledge develops on a basis of intellectual knowledge. Wisdom-related knowledge inductively reduces the quantity and complexity of intellectual knowledge in favor of what is deeper and more essential. Wisdom researcher Sternberg defines wisdom as follows:

“the application of tacit knowledge towards the application of a common good through a balance among intra-, inter-, and extra-personal interests to achieve a balance among adaptation to existing environments, shaping of existing environments, and a selection of new environments, over the long term as well as the short term.” (Sternberg, 1998)

One might summarize wisdom as “the ability to produce broadly beneficial intended results while taking the full consequences of behavior into account.” Again we find a combination of skill (tacit knowledge), and (implicitly) a pervasive (long term) understanding of reality, in combination with an urge to improve and shape the living environment. We consider this developing urge to improve and shape living environments an essential aspect of open-ended development and propose an explanation for that below in the section on a complex world.

AUTHORITARIANS AND LIBERTARIANS

The solid part of the spiral is the development up to the level of the authoritarian personality as defined by Stenner (2005, 2009). Authoritarians “*are not endeavoring to avoid complex thinking so much as a complex world*” (Stenner, 2009, p. 193). It is the authoritarian’s underdeveloped cognitive capacity that “*reduces one’s ability to deal with complexity*.” This personality-type seeks, appreciates, and even demands external authorities to maintain the living conditions in which they can function adequately: normalcy. For authoritarians “authorities” are the processes or agents that they perceive as responsible for maintaining normalcy (and with that their sense of adequacy). Authoritarians display “*bounded autonomy*” because they exhibit autonomy only in a suitably controlled environment. Authoritarians actively help their authorities in a particular and highly characteristic way: by reducing the perceived complexity of the environment; in particular through intolerance of diversity and by supporting some perceived central authority (an agent or process) with the same surmised aim.

The dashed part of the spiral progresses beyond this level to the libertarian personality (Stenner, 2005). Libertarians have gradually developed the autonomy and skills to co-create living conditions in which they and others feel and act adequately without the need for external authority to maintain and create these conditions. Libertarians have internalized the role of authority and prefer therefore individual authority to centralized authority. As such libertarians become local centers of development and growth in their (social) environment and consequently centers of diversity. Compared to authoritarians who can function adequately in standard situations and tend to exhibit norm-complying and norm-returning behavior, libertarians (have learned to) understand the world to a degree that they can cope effectively with deviations from normalcy and they use the benefits this provides to enhance their lives.

Stenner used a very simple “child-rearing values test” (Stenner, 2005) to determine whether individuals were authoritarian or libertarian (she only used the extremes in her analysis). Participants that clearly preferred children to be raised as obedient conformist were deemed authoritarian and those that preferred children to be raised as independent self-deciders were deemed libertarian.

Apparently this simple six two-option test was enough to separate people into a group that aims to avoid (a more) complex world and a group that can comfortably deal with some more complexity. Stenner specifically identifies the reaction to “normative threads,” perceptions of leadership failure and diversity in public opinion, as key difference between authoritarians and libertarians.

Authoritarian behavior depends on whether or not the situation might develop beyond coping capacity. This entails that “*individuals with a certain level of authoritarianism may manifest entirely different attitudes and behaviors from one occasion to the next, depending upon the presence or absence of normative threat*” (Stenner, 2009, p. 189). And “*normative threat only invites this kind of fear, cognitive unraveling and out-bursts of intolerance among authoritarians, whereas in fact these very same conditions (i.e., the public dissension and criticism of leaders that are the hallmarks of a healthy democracy) induce only greater tranquility, sharper cognition, and more vigilant defense of tolerance among libertarians*” (Stenner, 2009, p. 193). We will use this observation in the next section to differentiate between Two Attitudes Toward a Complex World.

MAIN AXES

The axis from the center leftward in **Figure 1** reflects increasingly more advanced responses to environmental challenges developing from innate (e.g., sucking), via emotional (e.g., happy or frustrated), to appropriate (e.g., culturally sanctioned) and even proactive responses (e.g., preventing future problems or creating a better society). Protruding downward is an axis denoting autonomy development. This axis develops from no autonomy at all, via the bounded autonomy of authoritarians, to the autonomy of libertarians. Extending rightward is an axis reflecting strategies developing from voluntary movements and direct perception-action relations, via coping strategies for the here and now, to advanced co-creating strategies that define and shape the environment (i.e., the agent as authority).

The axis extending from the center upwards reflects a development from a dependent self, to an immature and mature self. This development of the self has two separate but related facets: social and personal maturity. “*Social maturity is defined by measures of adaptation such as life satisfaction, environmental mastery, or positive social relations. Personal maturity, however, is indexed by openness to experience and indicators of personal wisdom such as personal growth and ego development*” (Staudinger and Glück, 2011, p. 213). Development of the self moves people increasingly away from egocentric, dependent, and self-centered modes of being (in **Figure 1** referred to as “immature self”), toward the capacity to take perspectives on the self and others, and to experience positive, helpful, responsible, and mutual interaction with others referred to as “mature self” (Richardson and Pasupathi, 2005, p. 145).

DIAGONAL AXES

The lower left diagonal in **Figure 1** simply reflects the development from a baby, which is preoccupied with discovering its body and its immediate environment, to childhood in which it is preoccupied with the exploration of the neighborhood and the

acquisition of habits, skills, and knowledge, and to adulthood in which one's potential can be developed and utilized in full.

The upper left diagonal shows the three main stages of moral development as described by Kohlberg (1971). Kohlberg calls the first main stage "pre-conventional" in which *the child only understands the consequences of its behavior in terms of direct effects on self in terms of (un)pleasantness and in which it knows that obedience is a way to avoid punishment*. At this stage *right action concerns mainly the satisfaction of one's needs*. In the second, "conventional," phase *the individual's attitude is not only one of conformity to personal expectations and social order, but of loyalty to it. It actively maintains, supports, and justifies the order and identifies with the persons or group involved in it*. This phase corresponds closely to the description of authoritarianism. The third stage is called the "post-conventional," "autonomous," or "principled" level. Individuals at this stage *make a clear effort to define moral values and principles that have validity and application apart from the authority of the groups of persons holding them and apart from the individual's own identification with the group* (Kohlberg, 1975). This stage corresponds closely to the description of libertarians. A 20-year longitudinal study in Chicago found moral judgment development to be positively correlated with age, socio-economic status, IQ, and education. In addition development in childhood predicted development in adulthood. At age 36 only about 10% had reached a moral development at post-conventional level (Colby et al., 1983); this suggests indeed that it is more an option than a default in modern Western cultures.

The upper right diagonal reflects words from the field of epistemological development [see van Rossum and Hamer (2010) for an overview] and in particular from Kuhn et al. (2000) who separates four levels of beliefs about the world. In the first "realist" level, assertions exist only in direct reference to a state of the world. In the second "absolutist" level assertions are authority derived true or false representations of the world. In the third level assertions are opinions that can be freely chosen, are accountable to their owners, and that, apart from authority support, cannot be ranked in terms of quality. In the fourth level assertions are judgments that can be evaluated and compared according to criteria of argument and evidence. This fourth level has passed what van Rossum and Hamer (2010), p. 26 call the watershed between reasoning in terms of ready-made things (facts, procedures) existing "out there" to independently constructing meaning. Since this is, again, a transition between dependence and independence of authority we associate (but not equate) the "watershed" with the transition from authoritarianism to libertarianism.

The last diagonal, in the lower right, describes typical activities associated with different life-phases. A baby is typically involved in all forms of sensory-motor explorations in which it gradually learns to separate the whole of perceptual and motor experiences into meaningful units. This parts-from-whole approach of participatory discovery is typically associated with the right brain hemisphere (McGilchrist, 2010). The second phase is typically culturally, technically, and representationally driven. In this phase the main sources of knowledge are represented and conveyed via languages (of diverse forms) and technologically and culturally constructed objects and environments. This is a

phase in which—in our Western cultures—the left hemisphere is dominant. It is also a world in which knowledge and skills are constructed from parts-to-whole. Knowledge and skills are typically not self-discovered but directly derived from others (authorities). In the post-watershed phase the participatory co-creation that characterizes self-actualized individuals takes again the effect of behavior in an ever-extending context into account. This suggests a return to right hemispheric dominance. The processes that drive these developments, and the rational to assign them to dominant hemispheres will be addressed in the next section.

The next section addresses two essentially distinct modes of being that are believed to underlie both hemispheric differences as well as the key properties of intrinsic and extrinsic motivations that, we claim, differ in the way they approach a complex situation.

TWO ATTITUDES TOWARD A COMPLEX WORLD

Complexity research has shown (Kauffman, 1995; Capra, 1997) that all life and therefore all human activity seems to occur in the transition region between order and disorder or structure and chaos (Mora and Bialek, 2011). Too much structure precludes diversity and development. Too much disorder precludes stability and predictability. Put differently: moderately increasing disorder allows for more diversity and development but allows less control. In moderation, disorder may lead to novelty, in excess it leads to chaos. In contrast, increasing order fosters uniformity, predictability, and control, but in excess it leads to stagnation and lifelessness. Note that the moment a novel structure has been discovered in a previously disordered or chaotic state, some order (and meaning) is imposed on it and the complex system becomes a little more tractable and accessible to agent influence. With this discovery the "edge of chaos" has been pushed toward higher complexity. We propose that this process pushes development along the spiral in accordance with Vygotsky (1978) zone of proximal development.

TWO MODES OF COGNITION

We can call the form of cognition that allows us to discover novel structure "cognition for disorder," "cognition for possibilities," or "explorative cognition." Whatever it is called, its essential nature is participatory: structures in (apparent) chaos are only discovered through some form of participation in the system. During exploration and play, the properties of these structures are revealed and the structures of interest become gradually more familiar and predictable. This allows their properties to be generalized, abstracted, and integrated with existing knowledge and in doing so made useful for in the widest possible range of environments and (individual) challenges.

In situations where errors are costly (or even deadly) we need a complementary form of cognition: a form that more aptly is called "cognition for order," "cognition for certainty," or "control cognition." Both are essential forms of cognition and together they allow for a gradual proven and reliable extension of the limits of agent capability toward ever more complex situations and ever-larger temporal and spatial scopes. This continual progression of exploration, consolidation, and testing is another formulation of open-ended development.

Recall that the reaction to an increasing complex world is the key difference between authoritarians and libertarians (Stenner, 2009). This suggests that the complexity of our (living) world is a deciding factor in determining whether someone is (or behaves as) authoritarian or libertarian. Authoritarians tend to abhor a complex world and feel an urge to reduce its complexity, while libertarians can deal comfortably with some additional complexity. The authoritarian reaction to increased complexity is with fear and intolerance of diversity (reducing complexity), while the libertarian reacts with increased interest and sharper cognition (mastering complexity). This suggests that explorative cognition and control cognition, in particular with authoritarians, are activated depending on whether the environment is appraised as safe or unsafe.

The depiction in **Figure 2** visualizes these two cognitive responses. The backdrop is Escher's, 1955 tessellation "Liberation" that reflects a progression from lifeless, predictive structure toward living free dynamics and endless possibilities. Here we assume that an agent's coping capacity allows it to deal with some intermediate level of complexity half way this progression. Depending on whether the overall situation is perceived as safe or unsafe, an agent might be motivated to explore dynamic diversity and novelty—the interest bias—or be motivated to reduce the complexity of the environment by helping to reduce the complexity through curtailing diversity and dynamics—the fear bias. The higher the life-fraction spent with an interest bias, the more one explored and the more one learned to master complexity.

It is therefore not surprising that the personality trait "openness to experience" correlates positively with libertarianism (Stenner, 2005). According to McCrae and Sutin (2009) "*highly open people are thus seen as imaginative, sensitive to art and beauty, emotionally differentiated, behaviorally flexible, intellectually curious, and liberal in values. Closed people are down-to-earth, uninterested in art, shallow in affect, set in their ways, lacking curiosity, and traditional in values.*" This contrast reads as a preference for an interesting vs. an ordered world. In addition "*open people admire openness, closed people despise it*" (McCrae and Sutin, 2009)." Associated with a closed attitude is "the need for closure" (Kruglanski and Webster, 1996; Malhotra et al., 2008), the desire for definite and final answers. People prone to seizing on the first idea offered and then freezing on this solution are in general uninterested in exploring alternative possibilities, keeping their views simple and uncluttered.

TWO HEMISPHERES

The existence and detailed properties of these two forms of cognition have recently been described in the seminal work on the divided brain by McGilchrist[2010; see Rowson and McGilchrist (2013) for a highly accessible introduction]. McGilchrist argues that the two cortical hemispheres understand the world in quite different ways. In particular it suggests to us that the left hemisphere specializes in cognition for order, while the right hemisphere specializes in cognition for disorder. **Table 1** provides a representative fraction (McGilchrist, 2010 chapter 1) of the wealth of reported differences

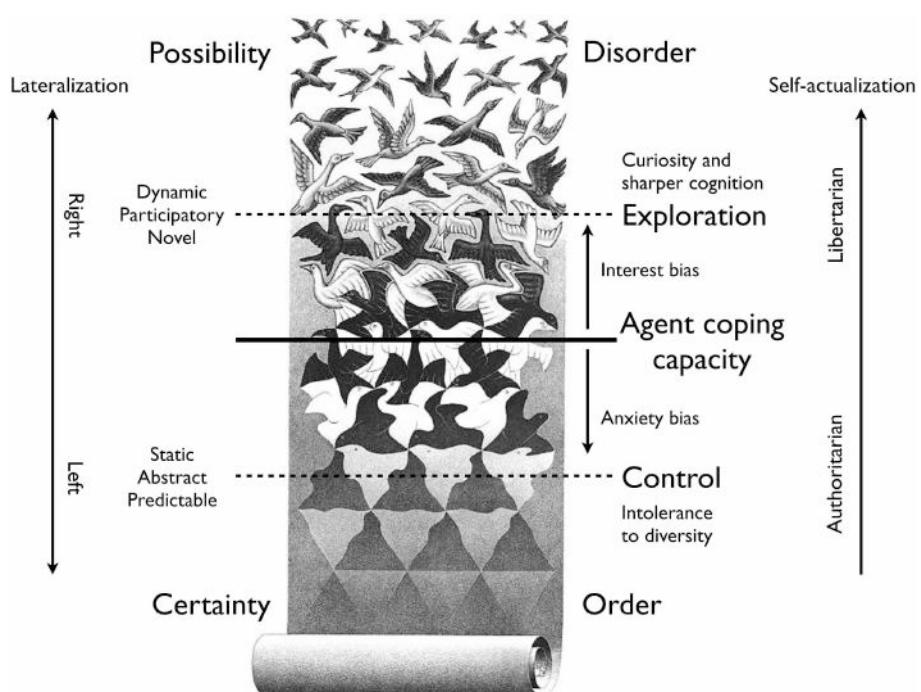


FIGURE 2 | Dealing with complexity. The anxiety-free response to increased complexity leads to curious exploration and sharper cognition, while the anxiety-laden response activates intolerance of diversity. This graphical depiction can be interpreted as agent development at some

part of the spiral in **Figure 1** that gradually moves outwards toward self-actualization. (M.C. Escher's "Liberation" © 2013 The M.C. Escher Company—the Netherlands. All rights reserved. Used by permission. www.mcescher.com).

Table 1 | Overview of roles and approaches ascribed to the left and right cortical hemisphere that together define two quite different stances toward the world.

Topic	Left hemisphere Cognition for: control, order, certainty	Right hemisphere Cognition for: exploration, disorder, possibility
Main requirements	Associated with fear and anxiety, detachment, abstract manipulation, closed to experience.	Associated with interest, participation, interaction, play, openness to experience.
Closed vs. open personality	Closed people are down-to-earth, uninterested in art, shallow in affect, set in their ways, lacking curiosity, and traditional in values. They are prone to seizing the first idea offered and stick to it to keep their views simple and uncluttered.	Openness to experience: imaginative, sensitive to art and beauty, emotionally differentiated, behaviorally flexible, intellectually curious, and liberal in values.
Main concern	Principal concern is utility: the world as a resource.	Prioritizes what actually <i>is</i> and what concerns us.
Scope	Local short term view. Deal with what it knows.	Bigger picture (broader, long-term view). Draws attention from the edges of awareness.
Interests	Interested in the familiar and the known, difficulty with disengaging from the familiar. Concerned with what it knows. Concerned with man-made objects. Non-living objects specialist. Living entities as tools or instruments. Body-parts. Tools and machines.	Interested in the novel. Concerned with what it experiences. New information, new skills, emotional engagement. More concerned with living individuals. Living individuals as other individuals. Food + musical instruments. Body as a whole.
Preferences	Preferences for things that are represented as relatively invariant across specific instances, allowing for abstracted types or classes of things.	Preference for things that exist in the world. Sensitive to what distinguished different instances of similar type from each other.
Strengths	Thoroughly known and familiar. Efficient in routine situations and familiar skills. Prioritizes the expected and generates expectations. Things made fixed and equivalent: types. All that is re-presented as over-familiar, inauthentic, lifeless [because not individuated] categories.	Gathering new information. Good when prediction is difficult. Anomaly (individuality) detector: individuals. More efficiently when initial assumptions need to be revised or when old information needs to be distinguished from new information. All that is "present" as new, authentic, and individuated.
Attention type	Local narrowly selective (highly) focused attention.	Broad, global, and flexible attention.
Attitude toward world	Representing the world: the world as a copy that exists in conceptual form, suitable for manipulation.	Experiencing the world: the world as it is, open for novelty and whatever exists apart from ourselves, without preconceptions and not focusing on what it already knows.
Construction of world	Start with pieces and put these together. Bottom-up.	Start from the whole and go, if required, into detail. Top-down.
Representation of objects	Preference to re-present categories of things, and generic, non-specific objects.	Individual unique instances of things and individual generic objects: individuals are <i>Gestalt</i> wholes. Concerned with the uniqueness and individuality of each existing thing or being.
Solution limitations	Problem solving: single solution and latch on to that. Deny inconsistencies. Suppressing not currently relevant relations.	Array of possible solutions, which remain live when alternatives are explored. Actively watching for discrepancies.
Associations	Single strong association more important than multiple weaker associations.	Widespread activation of relations. Single strong or multiple weaker relations equally important.
Preferred knowledge type	Affinity with public knowledge.	Personal knowledge.
Identification	Identification by parts. Gradual (knowledge-based) construction.	Identification from/by whole. "Aha!" phenomena through seeking and finding patterns in things.

(Continued)

Table 1 | Continued

Topic	Left hemisphere Cognition for: control, order, certainty	Right hemisphere Cognition for: exploration, disorder, possibility
Reasoning	Linear sequential arguments. More explicit reasoning. Concentrating helps to focus on explicit structure of the problem.	Deductions and some kinds of mathematical reasoning. Pleasurable "Aha!" phenomenon mediates between emotions and higher frontal cognitive functions. Insights when NOT concentrating on a problem. Link with anomaly (inconsistency detection in own assumptions). Concentrating on problem impairs finding a solution.
Language use	Language as symbol manipulation, More extensive vocabulary, subtle and complex syntax. Parsing of utterance, but meaning less deep. Explicit meaning.	Interpretation as a whole and in context, attribution of full meaning. Use of intonation and pragmatics. Non-literal and implicit meaning. Sensitive to subtle unconscious perception. Better at detection deceit.
View on world	More optimistic view of the self and the world. Also unwarranted optimism. More anger.	More associated with sadness than with anger. Sadness associated to low activation of frontal lobe.
Main emotions	Emotions associated with competition, rivalry, individual-self-believe (positive and negative).	All emotions. Emotions related to bonding and empathy.
Empathy	Unconcerned with others and their feelings.	Empathic identification. Self-awareness, empathy, identification with others. But only with what is known [considered] to be another living being—not a mechanism. Theory of mind.
Link with older parts of the brain and body		More connected to the limbic system and the ancient subcortical systems. Hypothalamic-pituitary axis, which is where the endocrine interfaces with body and emotion. Essential to the subjective appreciation of the body's physiological condition.

The description in the three header rows stems from the requirements of cognition for order and disorder. The header of the table summarizes cognition for order and cognition for disorder. The body of the table contains near literal formulations from chapter 1 of McGilchrist (2010).

in how the individual hemispheres approach and understand the world.

McGilchrist argues that in the last two or three millennia, our Western societies have become characterized by an ever growing dominance of the left-hemispheric world view that favors a narrow focus over the broader picture, specialists over generalists, fragmentation over unification, knowledge and intelligence over experience and wisdom, technical objects over living entities, control over growth and flourishing, and dependence over autonomy. In his book, called *The Master and His Emissary* McGilchrist argues that the right hemisphere, with its holistic perspective and more intimate relation with the body is the master that tasks its emissary, the left-hemisphere, with focused assignments. However, in our increasingly culturally defined (i.e., more technically structured and less naturally organized) world, where linguistically transmitted shared knowledge has become more important than individually acquired tacit knowledge, left hemispheric strengths seem to have become more beneficial for most of us than right hemispheric strengths.

IN- AND EXTERNAL AUTHORITY

However, and this is essential for our discourse, the left and right hemisphere require quite different conditions to function optimally. The right hemisphere assumes autonomous participation

in an open, dynamic, and infinite world of nested dynamical systems that form dynamically stable and continually evolving entities. In this mode of being, truth is defined as accordance with reality and is to be tested by acting out in the world; right-hemispheric knowledge and experiences are essentially subjective. As such this mode of being is particularly effective in situations where new aspects of the dynamics of the world are to be investigated to expand the thought-action repertoire (Fredrickson and Branigan, 2005) and where novel and creative solutions are appropriate.

In contrast, the left hemisphere assumes a closed, static, and finite world in which entities are symbolic, discrete and abstract and in which one is an "objective" observer instead of a participant. In this mode of being, truth is defined as the result of consistent reasoning and consensually agreed on linguistically shared and presented facts. This mode of being is particularly effective in situations in which problems have to be solved or addressed in a detached, rational, standardized, and communicable way. Scientific communication is a typical example of this. Because of this more narrow focus, left hemispheric strategies essentially depend on processes that create and maintain the required closed, static, and finite world: the normative order introduced earlier. We argue that *authorities*—defined as processes or agents that create, maintain, and influence the conditions in which agents

exist—fulfill this role. Adequate left hemispheric strategies, we propose, are only possible if either an internal authority, i.e., the right hemisphere, or external authorities ensure that conditions are maintained in which left hemispheric strategies are effective.

In particular we propose that the authoritarian mode of being corresponds to a left hemispheric dominance in combination with a need for external authorities to create and maintain the conditions in which a dominant left hemisphere can function adequately. Libertarianism corresponds to a right hemispheric dominance that is able to provide the proper conditions for left hemispheric functioning. This entails that the authoritarian agent, as the name suggests, is essentially dependent on *external* authorities, while the libertarian agent, again as the name suggests, is free from external authorities because the agent is able to self-maintain the conditions in which both modes of cognition contribute adequately. To put it bluntly, we argue that authoritarianism in adults is a sign of arrested development that limits individual autonomy growth to environments maintained by external authorities.

AUTONOMY IN TWO OR THREE STEPS

In terms of **Figure 1**, this can be described as an initial right hemisphere dominated inner-loop in which one learns to master the body through playful interaction with the world. The second loop is left hemisphere controlled because one learns from external authorities and through abstracted linguistically conveyed knowledge about the structures of the world. However the purpose of this phase is to learn how to make the mind a useful instrument. If this process succeeds, it allows one to effectively produce intended results in both culturally defined and natural worlds. As such it is a basis for confidence, further exploration, and gradually increasing autonomy through the ability to co-create ever more extended (both in place and in time) environments in which one can self-maintain the condition for adequate functioning. This describes the third (pre-conditioned) loop.

However, when an agent is unable to make the mind into a reliable instrument, the individual is frequently confronted with the inability to produce intended results. And because the left hemisphere is dominant in this phase, one responds in the complexity reducing control mode favored by authoritarians. It is interesting that “power” is defined as “*the ability to produce intended results*” (Russell, 1938). Earlier we summarized Sternberg’s definition of wisdom (Sternberg, 1998) as “*the ability to produce broadly beneficial intended results while taking the full consequences of behavior into account*.” This suggests defining raw power as “*the ability to produce intended results without necessarily taking the full consequences of behavior into account*.” Its is therefore not at all surprising that typical centralized authoritarian organizations such as bureaucracies, governments, large corporations, and the military are always associated with “power” and standardization.

Libertarians do not need the control over the environment provided by these centralist structures and they are, because they made their mind into a reliable tool, not obsessed with reaching intended results (they can do that more often than not). In contrast they are more interested in understanding the full consequences of behavior. This requires a participatory approach in which one learns to discover and predict the innate dynamics of

the social, cultural, and natural world without necessarily controlling or curtailing its diversity. On the contrary, working *with* the inherent dynamics of the world is a way to stabilize it (or not to disturb it). We refer to this creative process of moving with the dynamics of the social and natural world as “co-creation”: a product of open-ended development.

In the next section we will argue that external drivers of behavior (functioning as external authority) are associated with extrinsic motivation and left hemispheric strengths, while internal drivers of behavior are associated with intrinsic motivation and as such with learning to co-create and open-ended development. We will use the appraisal of the environment as the link between open-ended development, the two attitudes toward the world, and motivation.

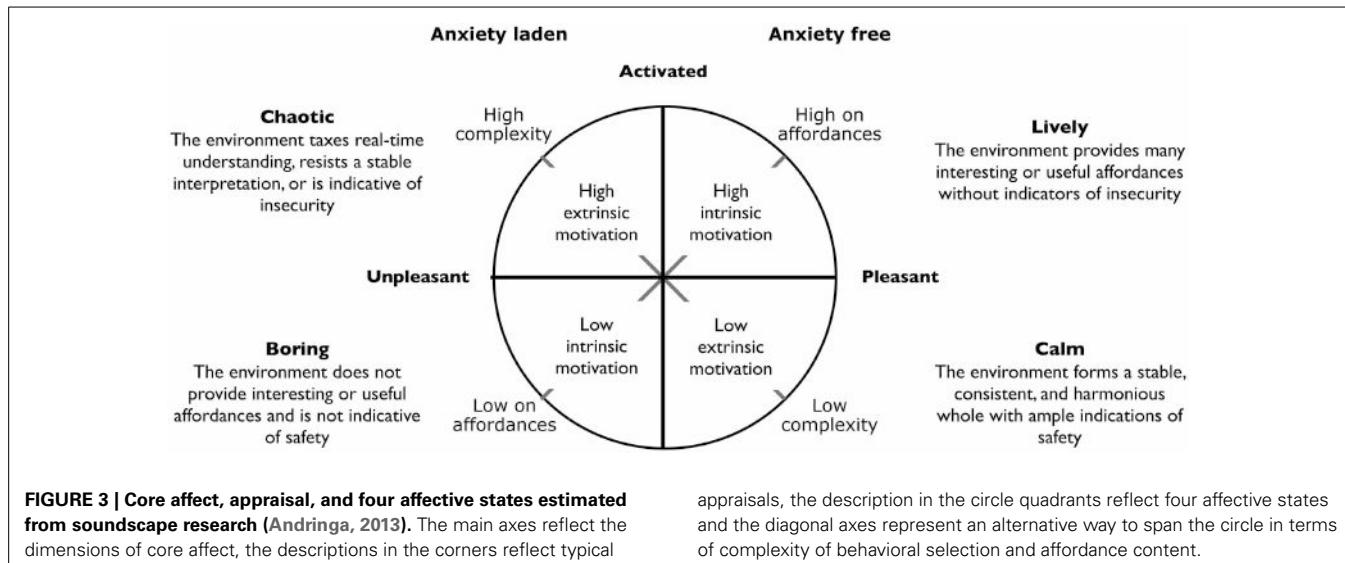
MOTIVATIONS

To be motivated means to be moved to do something (Ryan and Deci, 2000). However it is not yet clear *how* states of the world or states of the individual motivate agents to spend their (mind) time in particular ways. We will therefore start this section with some recent results from soundscape research that helped us to formalize the influence of the environment on motivation.

APPRAISAL, MOTIVATION, AND CORE AFFECT

A soundscape is a perceived sonic environment and soundscape research addresses the role of sounds and sonic environments on individuals and society. In a recent paper (Andringa and Langer, 2013), addressing how quiet sounds promote and annoying sounds impede health, we analyzed the words people use to appraise sonic environments (Axelsson et al., 2010). Appraisals are “*cognitive evaluations of events that are considered to be the proximal psychological determinants of emotional experience, with different combinations of appraisals corresponding to different emotions*” (Kuppens et al., 2012). Appraisals typically refer directly or indirectly to motivation. Kuppens et al. lists: motivational relevance (“*Is it important?*”); motivational congruence (“*Is it advantageous or disadvantageous?*”); agency (“*Is it caused by others or myself?*”); problem and emotion focused coping potential (“*Can I cope with the situation and with my emotions?*”); future expectancy (“*Is the expected outcome desired or not?*”). Appraising the environment therefore combines motivation, coping capacity, and expectations of the future. As such the appraisal process involves the evaluation of possible (inter)actions with the environment.

Appraisals are also connected to a central concept in emotion theory called “core affect” (Russell, 2003). Core affect is defined as an integral blend of the dimensions displeasure-pleasure (valence) and passive-active (arousal). Unlike emotional episodes, which are relatively infrequent, core affect is continually present to self-report. Core affect is usually visualized as a circle with the pleasure axis horizontally and the arousal axis vertically as depicted in **Figure 3**. Here relaxed and invigorated moods are situated in the lower and upper right quadrants and moods like boredom and anxiousness in the lower and upper left quadrants respectively. Associated with these moods are calm and lively appraisals on the right and a chaotic and boring appraisals on the left (Andringa, 2013). Appraisals and core affect mutually influence each other (Kuppens et al., 2012).



AFFORDANCES AND COMPLEXITY

Since appraisals involve the evaluation of possible interactions with the environment, they pertain to two main questions: what action opportunities does the environment afford and: how to decide on the best course of action? We will refer to the first question as the affordance content, in Figure 3 as the diagonal from lower left to upper right, and to the second as the complexity of the environment. We will address these issues in order. Because the description of the visual scene leads to quite similar patterns of descriptive words (Russell et al., 1981; Axelsson et al., 2010) we treat our results as if they pertain to perception in general.

Affordances are perceivable action possibilities, provided by an environment (Chemero, 2003) that might be used to satisfy (immediate or future) needs. Affordances arise thus from the interaction of the environment with the perception capabilities of the individual agent. Interesting environments provide discoverable affordances to extend knowledge and skills through, typically, playful interaction (Fredrickson, 1998). Boring environments are devoid of discoverable affordances and do not provide appreciated novelty (e.g., because they are devoid of stimuli, or the stimuli are either too static to be useful or too complex to interpret). The more one interacts (plays) with interesting environments the more complex affordances one learns to perceive.

The complexity of an environment is, in this context, a reference to how difficult it is to cope with environmental challenges and opportunities. Complexity therefore refers not to the environment *per se*, but to the question of how difficult it is for an agent to decide on situationally appropriate behavior. Low complexity environments are highly redundant (each part “predicts” the whole, leading to an impression of harmony), which entails that most perceptual evaluations of the environment lead to a similar overall interpretation of pervasive safety. In “calm” low complexity environments action outcomes are relatively insensitive to the details of action selection and action execution; one is neither forced nor enticed to act overtly and the mind is free to wander and to attend its own business (Andringa and Lanser, 2013).

In contrast, highly complex environments are less redundant; for example because of a lack of internal coherence due to a multitude of uncorrelated processes, giving an impression of chaos and unpredictability. This entails that the focus of attention needs to be chosen and adapted well to ensure a proper selection and execution of coping behavior. In contrast to low complexity environments, complex situations may force one to act in a highly controlled fashion and in response to particular events. This entails that action outcomes in complex environments are highly sensitive to detail.

This analysis suggests four qualitatively different types of (sonic) environments in terms of the complexity of action selection and affordance content. The complexity depends on the agent's ability to select a safe course of action. Highly complex or chaotic environments are difficult to interpret (e.g., due to an overabundance of diverse stimuli), actively indicative of insecurity, or in other ways requiring a precise selection of activities. This type of environment activates highly focused mind-states aimed at coping with the here and now. A boring (sonic) environment is low on useful (audible) affordances and is, for that reason, not indicative of safety, which activates alert mind-states. In contrast, a lively environment is not indicative of insecurity and represents many affordances that provide ample interesting opportunities to attend, and it allows one freedom to address the available affordances at will. The fourth environment is calm or relaxing because it provides ample indications of safety and allows as such full freedom of mind-states to relax and recuperate. Figure 3 provides these four domains of environmental appraisal.

In terms of the spiraling open-ended development depicted in Figure 1, the growing ability to detect and effectively use affordances is a measure of progress along the spiral. Initially the affordance content is predominantly used to determine situationally appropriate conformist behavior, but gradually the affordances can be used in the more individualized and situationally appropriate fashion characteristic of co-creation. Similarly, any growth of the agents coping ability in Figure 2 depends on

an increasing ability to perceive more and more complex affordances and to learn more and more generic and reliable coping strategies.

MOTIVATION AND REWARD SIGNALS

According to Baldassarre's (2011) recent paper, motivations are based on mechanisms that "drive learning of skills and knowledge, and the exploitation and energisation of behaviors." But extrinsic motivations do this "on the basis of the levels and variations of homeostatic needs detected within the visceral body," while intrinsic motivations "facilitate this, on the basis of the levels and the variations of such skills and knowledge directly detected within the brain." This suggests that, according to Baldassarre, motivations are exclusively based on information derived from either the body or the brain: appraisal of the environment plays no (explicit) role. In addition, skills and knowledge derived from extrinsic motivations "have the adaptive function to produce behaviours that allow the regulation of those homeostatic needs so as to increase fitness." In contrast "intrinsic motivations have the adaptive function to allow organisms to learn skills and knowledge without the necessity to have a direct impact on homeostatic needs and fitness at the time of the acquisition. These skills and knowledge contribute to increase fitness as they can later be used to learn, relatively quickly, complex behaviours and long chains of actions that regulate homeostatic needs."

Strictly interpreted this entails that extrinsically motivated behavior only occurs after the visceral body develops a homeostatic need, while intrinsically motivated behavior has no direct benefit. Consequently a well-fed agent on the track of an approaching train might be fascinated by the complex behaviors and long chains of actions afforded by this experience, but it will not move unless it timely develops a visceral need such as thirst. Yet apart from the absent role of situational appraisal there is much to agree with in Baldassarre's definition. In particular the role of the perceived needs of the visceral body—now or in the foreseeable future—that define extrinsic motivations and its reward function.

Ultimately, extrinsic motivations are deficiency motivations and are associated with what Maslow referred to as D-cognition ($D = \text{deficiency}$) which he defined as "*the cognitions that are organized from the point of view of basic needs or deficiency-needs and their gratification and frustration* {Maslow:1962tn, p. 189}." The reward signal of D-cognition is need-gratification: the pleasures of food after abstention, restoring order after chaos, relief after a negotiating a dangerous situation, or a monetary reward after boring work. Intrinsic motivations are uncoupled from direct need gratification and allow what Maslow referred to as B-cognition ($B = \text{being}$), a form of cognition in which the world (or objects as Maslow referred to) as it objectively exists can be discovered. These two forms of cognition again refer to the two modes of being outlined in section Two Attitudes Toward a Complex World. It is therefore to be expected that extrinsic motivations are predominantly left hemispheric phenomena, that are driven by utility, while intrinsic motivations are more right hemispheric phenomena associated with exploration and open-ended-learning.

Baldassarre (2011) details how intrinsic motivations provide the reward signals required to drive reinforcement learning. According to him "*intrinsic motivations are based on mechanisms that measure the success of the acquisition of skills and knowledge directly within the brain. For example, these mechanisms drive organisms to continue to engage in a certain activity if their competence in achieving some interesting outcomes is improving, or if their capacity to predict, abstract, or recognise percepts is not yet good or is improving: the brain detects all these conditions without involving the visceral body.*" The mechanisms that measure the successful acquisition of new knowledge, skills, and insights are essentially associated with open-ended development. The experience of this success has been described by Maslow (1954) as a feature of B-cognition. Maslow describes peak experiences as "*feelings of limitless horizons opening up to the vision, the feeling of being simultaneously more powerful and also more helpless than one ever was before, the feeling of great ecstasy and wonder and awe, the loss of placing in time and space with, finally, the conviction that something extremely important and valuable had happened, so that the subject is to some extent transformed and strengthened even in daily life by such experiences.*" According to Maslow the further the development toward self-actualization the more frequent these peak experiences occur, which suggests that they are experienced rewards signals that drive the later stages of open-ended development in B-cognition.

MOTIVATION, AGENCY, AND MIND-STATES

Motivation researchers such as Ryan and Connell (Ryan and Connell, 1989) couple motivations directly to the perceived locus of causality (PLOC), which reflects the degree the individual or some external authority or influence originates the behavior. It is a measure of autonomy and agency. The more autonomous the behavior, the more it is endorsed by the whole self and is experienced as action for which one is responsible (Deci and Ryan, 1987). This leads to a sequence of progressively more agentic motivations: "external," "introjected," "identified," and "intrinsic" reasons to act. According to Ryan (Ryan and Connell, 1989) "*external reasons were those where behavior is explained by reference to external authority, fear of punishment, or rule compliance.*" Introjected reasons are framed in terms of "*internal, esteem-based pressures to act, such as avoidance of guilt and shame or concerns about self and other-approval.*" These are typically situation-enforced motivations with the aim to prevent a worse outcome associated with doing nothing. "*Identifications were captured by reasons involving acting from one's own values or goals, and typically took the form of 'I want.'*" Through this identification the locus of causality shifts more and more to the agent. Intrinsic reasons for action occur whenever "*the behavior is done 'simply' for its inherent enjoyment or for fun.*"

More recently (Malhotra et al., 2008) ordered motivations in terms of intrinsic and extrinsic motivations that have an external or internal perceived locus of causality, and exogenous and endogenous motivation that reflect whether the behavior is driven either by external stimuli or by internal needs or drives. This resulted in four combinations of in-/extrinsic and exo-/endogenous motivations that dovetails very well with the four quadrants in **Figure 3** (combining appraisal and core affect), the

two modes of cognition in section Two Modes of Cognition, and the role of the two hemispheres described in section Two Hemispheres. As such this allows us to combine many concepts addressed in this paper in a single framework, which is depicted in **Table 2**.

The entries reflect descriptive words originating from different authors. The upper row and leftmost column reflect descriptions that pertain to the whole row or column respectively. The two rightmost columns, titled extrinsic and intrinsic, reflect modes of being that are directly associated with the two ways to approach complexity, the role of the left and right hemisphere, Maslow's D- and B-cognition, the role of safety in environmental appraisal, and the diverse descriptions of ex- and intrinsic motivations. The two lower rows reflect whether behavior is exogenous and highly activated or endogenous and less activated. The four remaining cells reflect descriptions that pertain to each of the different combination of in-/extrinsic and exo-/endogenous motivation. They also refer to a more general interpretation of the quadrants as depicted in **Figure 3**. These cells/quadrants have a descriptive name in bold.

The control quadrant reflects a combination of external motivating stimuli with the external perceived locus of causality characteristic of a challenging world. This quadrant reflects a motivational state in which an agent primarily aims to avoid immediate or future injury, harm, or disadvantage. Another name for this quadrant would be the problem-solving quadrant. An agent in this highly complex situation (in terms of behavior selection) is interested in any utility instrumental to avoid negative consequences and to retain or regain control. The associated mind-state

is stably focused on the problem as long as the problem exists and is a form of prolonged effortful directed attention (Kaplan, 1995).

The exploration quadrant combines external stimuli with an internal PLOC leading to self-chosen overt behavior that is perceived as fun and enjoyed for its own sake; all characteristic of an interesting world. Aimless but definitely unforced exploration and creation is only possible in apparent safety and requires environmental affordances at a level of complexity that the agent can handle without being taxed too much or too little. The associated mind-state is flexibly focusing on the most interesting aspects of the world, while remaining completely absorbed without lapses and pauses. Flow (Nakamura and Csikszentmihalyi, 2002) is a fitting description for this pleasurable mind-state.

The consolidation quadrant combines individual-need-driven activities with an internal PLOC. This is also only possible in a safe world. This may or may not lead to overt behavior, but is in all situations aimed at unforced self-development, growth, or other forms of psychological and physical recuperation and development. In this quadrant the associated mental activities are free to digress or to wander aimlessly without purpose or goal. One associated mind-state is fascination (Kaplan, 1995) which allows a prolonged, uninterrupted, and effortless immersion in an environment that is pleasant and self-selected to address personal needs proactively. This does not involve directed attention and therefore restores the capacity for directed attention. It is in this mind-state that the mind/brain can address its own needs.

The last quadrant is described with the term submission (to external forces), characteristic of a dominating world. This quadrant is characterized by an external locus of perceived causality

Table 2 | Four motivational states.

Motivations	Extrinsic	Intrinsic
Exogenous	Russell: unpleasant Ryan: external PLOC, low autonomy Maslow: D-cognition McGilchrist: left-hemisphere Baldasare: extrinsic, deficiency driven, direct fitness benefit Andringa: no safety, reactive	Russell: pleasant Ryan: internal PLOC, higher autonomy Maslow: B-cognition McGilchrist: right-hemisphere Baldasare: intrinsic, future fitness benefit Andringa: safety, pro-active
Control	World: challenging Ryan: introjected motivation (internal or esteem-based pressures to avoid harm) Malhotra: usefulness/utility Andringa: retaining or regaining control Andringa: high complexity Mind-state: directed attention	World: interesting Ryan: intrinsic motivation, completely self-determined activity Malhotra: hedonistic (fun, enjoyment) Andringa: learning and playing in safety Andringa: high affordances Mind-state: flow
Endogenous	Submission	Exploration
Russell: minimally activated Malhotra: Driven by internal needs/drives	World: dominating Ryan: external (authority enforced, fear of punishment, rule compliance) Malhotra: guided (to external regulation) Andringa: no sense of safety or control Andringa: low affordances Mind-state: boredom	World: safe Ryan: identified (personal importances) or integrated (personal goals) Malhotra: self-development, self-enhancement, self-growth Andringa: restoring resources and caring Andringa: low complexity Mind-state: fascination
		Consolidation

This table combines results and concepts from many different domains and provides a generalization of the quadrants in **Figure 3**.

in combination with unfulfilled internal needs that offer no other options than to accept guidance, to be subjected to external control (through threat, punishment, or fear), or to do nothing due to cognitive inadequacy given the current environment. In this quadrant the mind is never at rest, but fruitlessly in search of ways to cope. One associated mind-state is boredom, which is described (Martin et al., 2006) as “*Not being in control of life; agitated, yet at the same time, lethargic.*” In addition boredom is associated with restlessness, stress, the feeling of being trapped, frustration, fatigue, lack of concentration, guilt, meaninglessness, and even depression.

The range of scientific domains that have contributed to **Table 2** is wide and includes emotion research (Russell, 2003), motivation research (Ryan and Connell, 1989), human machine interfacing (Malhotra et al., 2008), computational development and learning (Baldassarre, 2011), soundscape research (Andringa and Lanser, 2013), personal development (Maslow, 1962), cognitive psychology (Kaplan, 1995), and general cognitive science and culture studies (McGilchrist, 2010). This is an impressive range that is suggestive of the fundamental nature of the topic of this call on open-ended development driven by intrinsic motivations.

OPEN-ENDED DEVELOPMENT DRIVEN BY INTRINSIC MOTIVATION

This concluding section returns to the core topic of the call: open-ended development driven by intrinsic motivation. We will use the four motivational states as described in **Table 2** to couple motivation to open-ended development via what we call the “open-ended development loop.” We will first address motivation in terms of attitudes and strategies to deal with the world as it is experienced. Secondly we will directly address the intimate relation between open-ended development, intrinsic motivation, and acting out in the world. Thirdly we will outline some consequences for artificial cognitive system research and in particular how to facilitate development toward truly autonomous and moral agents. Finally, we will argue that the left hemispheric biases characteristic of Western cultures have limited artificial cognitive systems research and we suggest a solution to address these limitations.

MOTIVATION, AUTHORITY, AND CO-CREATION

This subsection returns to the concepts “authority” and “co-creation” that we introduced as essential for open-ended development. We aim to demonstrate that they are important not only as core concepts of cognitive science, but also as defining concepts for agency and even as main forces that shape our (geo)political world.

The section End-State: Self-Actualization and Wisdom, discussed the target of open-ended development and concluded that the authoritarian personality type “seeks, appreciates, and even demands external authorities to maintain the living conditions (the normative order) in which they can function adequately.” In Section In- and External Authority we proposed that the need for external authority was a necessary consequence of left hemispheric dominance that requires a closed, static, and finite world to be effective. This entails that left hemispheric strategies

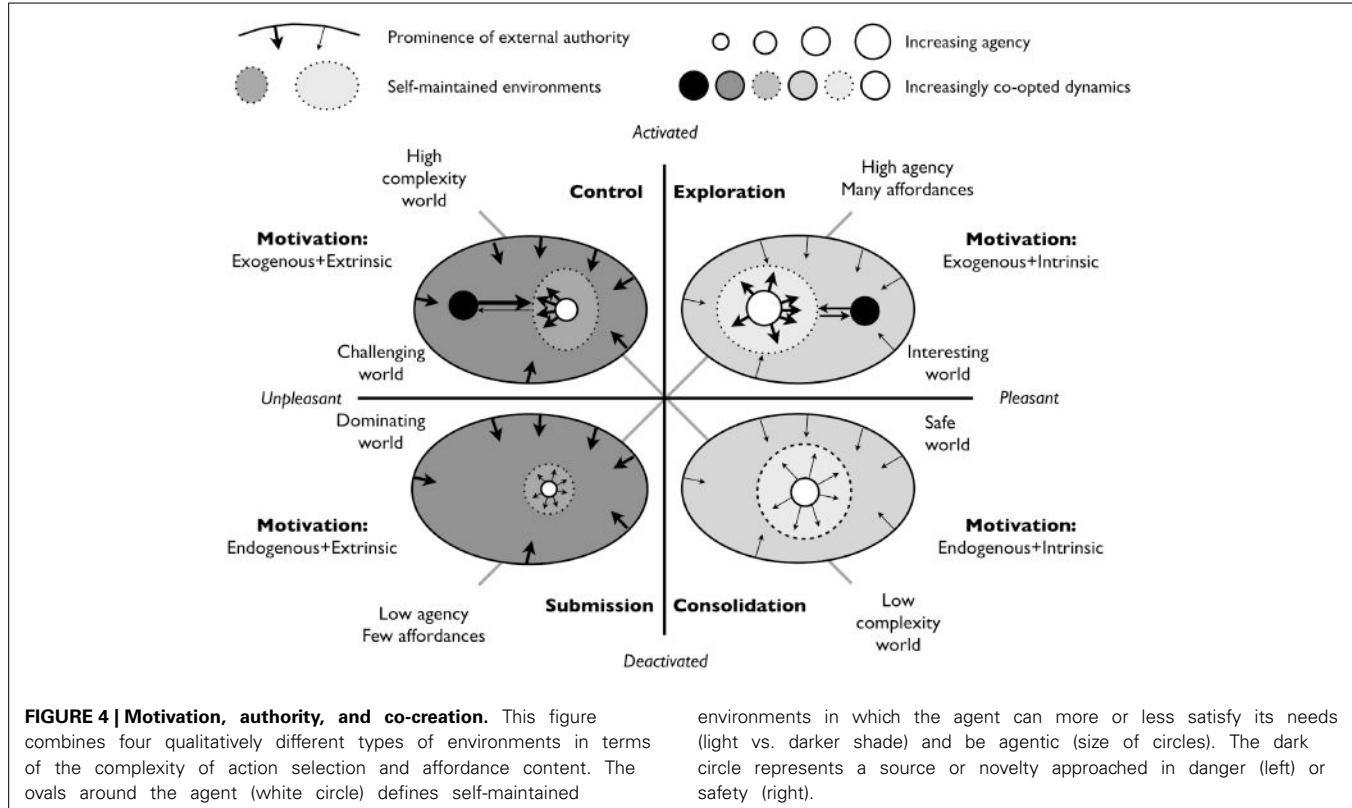
and external authority are mutually dependent: external authorities are expected to maintain the conditions in which the left hemisphere functions adequately. Left hemispheric strategies—through for example intolerance of diversity—reinforce the impact of external authorities through actively allowing external authorities more control while reducing one’s own agency. Overall this mode of being reduces the complexity of the world through increased uniformity and shared or centralized authority: the defining characteristics of authoritarians (Stenner, 2005). In moderation this process accounts for the existence of corporations, governments, organized religion, and the military. In excess it leads to stultifying bureaucracies in each of these organizations and eventually to oppressive dictatorship.

However, control through increased uniformity and centralized authority is neither the only nor the best way to deal with a complex world. Section End-State: Self-Actualization and Wisdom concluded that self-actualized or wise individuals feel a moral obligation to contribute to an improved world and we summarized wisdom as “the ability to produce broadly beneficial intended results while taking the full consequences of behavior into account.” Section Authoritarians and Libertarians concluded that the libertarian personality developed the autonomy and skills to co-create living conditions in which (s)he and others feel and act adequately, without the need for external authority to maintain and create these. The driving dynamics for this ability is rooted in the self-confidence resulting from the interest-based exploration and playful behavior that prepared the agent well for an unknown future (Silvia, 2008). In effect this leads to the ability to co-create ever more extended (both in place and in time) environments in which one can self-maintain the condition for adequate functioning, which leads to increasing diversity and individual authority: the defining characteristics of libertarians (Stenner, 2005). This advanced ability characterizes the outer (pre-conditioned) loop of the spiral development in **Figure 1**.

The difference between the authoritarian and libertarian mode of dealing with a complex world can be (Horney, 1945) summarized as “moving against” or “moving with.” The (according to Horney pathological) “moving against” mode controls diversity and reduces complexity through actively suppressing the inherent dynamics of the world. Note that this is the defining characteristic of our psychology or robotics labs. The (non pathological) “moving with” works *with* or co-opts the inherent dynamics of the world to stabilize it or to prevent the disruption of reliable and useful inherent dynamics. As such “moving with” is a summary of right hemispheric strategies.

The “moving with” mode of being, characteristic of the wise and the self-actualized, allows them not only to create and maintain an individual environment in which they can function adequately, it allows them to co-create the wider environment by gradually reducing the need for external authority (also in others) by (re)allowing and shaping the inherent dynamics of the world in favor of all its inhabitants.

Figure 4 provides a graphical depiction of much of the information in **Table 2**, but it focuses on the relation between the agent and the environment and the difference between external (controlling) authority and internal (co-creating) authority. The large ovals reflect the agent’s world that is more (light gray) or less (dark



gray) congruent with agentic needs. The prominence of external authorities (the inward pointing arrows) determines whether the world is characterized by suppressed dynamics (the authoritarian mode on the left) or co-opted dynamics (the libertarian mode on the right). The more the agent is able to create and extend a stable agent-maintained environment (dashed oval), the safer and more authoritative it is.

Figure 4 provides the four motivational quadrants defined in terms of the quadrants shown in **Figure 3** and **Table 2**. In the left quadrants the agent is either trying to control or is actually controlled by complex and ill-understood external forces that function as authorities. In the upper left quadrant the agent is challenged by environmental and/or agentic influences which stretch its coping capacity, force it into a narrow range of coping behaviors, and depletes its resources. In the lower left quadrant the agent is part of a world that is mainly beyond its control and understanding, since it does neither afford the agent useful affordances nor resupply of resources. As such it has to accept a minimally agentic role, for example by being forced to participate in activities that may harm its future interests.

In the quadrants on the right the agent's world is congruent with its needs (the most prominent of these is safety). The agent in the upper right quadrant is maximally agentic since it is able to use and explore the affordances of its world in safety and with satisfied basic needs. The agent exists in an interesting world in which it is free to participate in co-creation strategies that gradually elucidates and stabilizes more and more of the world's inherent dynamics for shared benefit. The agent in the lower right quadrant exists in a safe, low complexity environment. It is

unforced since, in essence, it profits from earlier co-creation activities of itself and others. This state allows the agent to resupply its resources (to address its needs) and to consolidate its experiences into generalized knowledge and skills.

This then, we conjecture, defines the success of open-ended development: successful open-ended development is characterized by a balance between the co-creation of a low complexity world, in which behavior selection is easy, in combination with high agency due to an abundance of affordances for maintained and extended co-creation. It is this dynamic balance that living agents find highly pleasurable. The enjoyment of successful agentic life—happiness—is therefore deeply meaningful: it is body and mind agreeing on success. And it also suggests that strengths of the right hemisphere, as listed in **Table 1**, might be understood as pervasive optimization.

OPEN-ENDED DEVELOPMENT DRIVEN BY INTRINSIC MOTIVATION

In this subsection we will more directly address the intimate relation between open-ended development, intrinsic motivation, and acting out in the world. In their review paper on extrinsic and intrinsic motivations and their importance for education and development, Ryan and Deci (2000) conclude that “*social contextual conditions that support one's feelings of competence, autonomy, and relatedness are the basis for one maintaining intrinsic motivation.*” They define relatedness as the basic need to feel connected, competence as the basic need to be effective, and autonomy as the basic need to feel agentic. According to Ryan and Deci we need these three basic human needs to be fulfilled in the classroom “*as one is exposed to new ideas and exercises new skills.*”

Interestingly, this conclusion can be connected one-to-one with the quadrant structure of **Figure 2**, **Table 2**, and **Figure 4**. In the exploration quadrant one expresses autonomy and agency and extends one's behavioral repertoire. In the consolidation quadrant one develops—in the absence of environmental pressures—new connections between oneself and the environment and one relates and combines hitherto unrelated knowledge and experiences. In doing so one generalizes, stabilizes, and consolidates knowledge and relations (whether mental, social, or otherwise). The consolidated knowledge, (social) relations, and skills, no longer new and unpredictable, become more and more suitable for general utility and in particular problem solving (a left-hemispheric activity). This corresponds to the problem-solving quadrant in which the agent can prove its increased competence and test and fine-tune its extended behavioral repertoire. Successful real-world problem solving leads to confidence, which is a basis for further exploration, consolidation, and testing. This “open-ended development loop” is depicted in **Figure 5**.

The continuation of the open-ended development loop depends crucially on the success-rate of the in the real-world problem solving ability. Failure to come up with a suitable solution leads to reduced confidence and eventually frustration. Perkins and Hill (1985) provide strong support that boredom is associated with frustration, and since the lower left quadrant is associated with boredom, low agency, and the need for guidance, it makes sense to situate persistent failure and the ensuing low confidence and reduced urge to explore in this quadrant. Persistent failure not only disrupts the open-ended development loop, it is also a strong demotivation to engage in any agentic activity and especially activities that are not habitual (because habits are activated by the environment) and therefore rely on some measure of agency.

This description is reminiscent of the phenomenon of learned helplessness that was discovered when “dogs exposed to inescapable

and unavoidable electric shocks in one situation later failed to learn to escape shock in a different situation where escape was possible (Maier and Seligman, 1976).” Learned helplessness depends on the uncontrollability of the aversive stimulus, which may entail that the agent learns that its activities do no longer produce intended outcomes. If so the agent does not unlearn its behavior, it simply no longer activates it because of its expected futility. Interestingly, in rats learned helplessness occurs only when one crucial condition is satisfied: “the response used in the test for learned helplessness must be difficult, and not something the rat does very readily.” Which, indeed, suggests that learned helplessness occurs only with activities that are agentic. This is the reason why the lower left describes its effect as “deactivating behaviors.”

RELEVANCE TO COGNITIVE SYSTEM RESEARCH

We believe that for autonomy to arise in any meaningful way, goal selection and achievement must occur in a (broad) transition region between order and disorder in which both danger and opportunity and defined (conform **Figure 2**). Without access to such a transition region and the experiences that it affords, the flexible and opportunistic balance and the complementarity between cognition for order en cognition for disorder cannot develop, which entails that there is nothing to drive the open-ended development loop in **Figure 5**.

Figure 5 suggests a principled way to formulate and structure reward signals because each of the quadrants may be associated with particular reward signals: the lower left with the gratitude of being led or adoration of authority, the upper left with the joy of restoring order, solving problems, or to receive social esteem rewards, the lower right with the joy of insights and understanding and the joys of interpersonal relations (love, friendships, and altruism), and finally the upper left the joy of play, exploration, and creation. The varying states of the environment and the associated appraisals (interesting, safe, or challenging) then

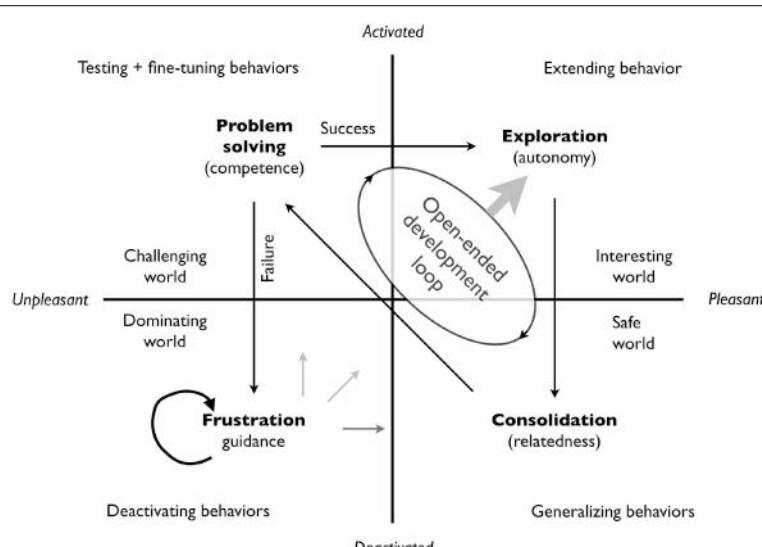


FIGURE 5 | Open-ended development loop. The words in brackets originate from Ryan and Deci (2000). The loop depends essentially on the rewards signals associated with exploration (experiencing novelty), consolidation

(discovering and fostering relations), and successful problem solving. The reward signals associated with this loop, described as peak experiences (Maslow, 1962), drive the outward spiraling development of **Figure 1**.

bring one in different learning modes. A suitable artificial agent that can engage in this open-ended development loop should be able to learn its way from guided exploration, consolidation, and problem solving into gradually more autonomous exploration, consolidation, and problem solving. In theory, each agent can learn to become autonomous and even wise (i.e., effectively co-creative), as long as it exists in an open environment that offers opportunities for all reward signals.

By constraining the learning environment it is possible to define the character of this agent by ensuring that it is not sufficiently exposed to all reward signals of the open-end development loop. For example, by making it very difficult to continue open-ended learning beyond a certain level of bounded autonomy, one creates an agent who will predominantly experience the reward signals associated with the pleasure of being “moved by.” This will lead to an agent that seeks and loves its servitude.

Alternatively an agent that is “raised” in insecurity will be exposed predominantly to the reward signals and learning outcomes associated with “moving against” uncertainty (e.g., of suppressing diversity), successful problem solving, protocol following, and other forms of cognition for order. This agent will be a quite autonomous apparatchik, someone *“not of grand plans, but of a hundred carefully executed details* (Billington, 1980), who has no inkling of its role in the grander scheme of things, and who will spontaneously and quite ruthlessly seek, accept, and support external authorities to maintain or restore the conditions for its adequate functioning. Characteristically It will enforce global uniformity and suppress local optimization whenever it increases diversity.

The agent that has been raised in a safe and protected situation and has primarily been exposed to the reward signals associated with love, friendship, and understanding, will develop the many facets of relatedness and profound interest in the world conform the induction capacity of cognition for disorder. This empathic agent will “move toward” others, be comfortable with diversity, and quite able to perceive and understand the beauty and ills of the world. However in times of adversity the empathic agent will not be able to organize or restore and maintain order the way the apparatchik can and it will probably be crushed by its imposed order and intolerance to diversity.

Fourthly the explorative agent is always in search of the reward signals associated with discovery, novelty, creation, and individual expression. This might be an artist agent who seeks the most individual expression of the most individual emotion, a risk-taker, or an autarchic agent that prefers the solitude of self-sufficiency to celebrate its individuality and autonomy. In Horney’s (1945) terminology he is “moving away.”

It is interesting that Horney’s (1945) terminology—moving away, moving toward, and moving against—fits so well on the three quadrants that define the open-ended development loop. Horney’s “moving with” personality, who moves with the dynamics of the world, is her only non-pathological personality. In our framework this is the personality that has learned from all reward signals and that as such has spent much time in the open-ended development loop, with the associated peak experiences. This is the only agent personality who has a proven competence (and autonomy) for most of its existence.

INCLUDING THE RIGHT HEMISPHERE IN COGNITIVE SYSTEMS

RESEARCH

Gomila’s and Müller’s (2012) definition of an cognitive system as “one that learns from individual experience and uses this knowledge in a flexible manner to achieve its goals” dovetails with how we defined raw power in section Autonomy in Two or Three Steps: “the ability to produce intended results without necessarily taking the full consequences of behavior into account.” In that section we concluded that executing raw power is a typical left hemispheric (authoritarian) response. A more developed libertarian, and wiser, response takes the full consequences of behavior into account. This suggests that the left hemispheric dominance of Western societies that McGilchrist (2010) describes has also limited the understanding of the artificial cognitive systems community by focusing its research on left hemispheric strongpoints such as object manipulation, problem solving, and task execution. If so, these Western biases have prevented the artificial cognitive systems community (and other scientific communities) from fully realizing the importance of right hemispheric strengths.

It might therefore be useful to study cultures without these Western limitations. For example Erica Fox Brindley, who studies the intellectual and cultural history of early China (500 BC to 200 AD), wrote a book on individualism in early China [for a summary see Brindley (2011)], which provides a rich description of the roles of agency, autonomy, and authority as the right hemisphere might understand these. She writes for example (Brindley, 2010 pp. xxvii–xxviii):

Earlier Chinese forms of individualism do not generally focus on the radical autonomy of the individual, but rather on the holistic integration of the empowered individual with forces and authorities in his or her surroundings (family, society, and cosmos). For early Chinese thinkers, there is no such thing as unfettered autonomy or freedom of will, in line with Kantian notions of the self. While such concepts are considered problematic even in some Western traditions they nonetheless constitute a core strand of thought that continues to inform contemporary concepts of individualism. In contrast to such conceptualizations, there exists a relative and relational sort of autonomy in early Chinese contexts, a type of autonomy that grants individuals the freedom to make decisions for themselves and to shape the course of their own lives to the fullest degree that they can and should—all from within a complicated and rich system of interrelationships. This type of autonomy, in other words, grants authority to the individual to fulfill his or her potential as an “integrated individual.” The goal of such an individual is to achieve authoritativeness as a person while at the same time conforming to certain types of authority stemming from his or her larger environment.

... Yet the emphasis in the Chines tradition on the relative autonomy of an individual from within a system of holistic and interconnected processes is quite different from many of the models with which we [Westerners] are most familiar. Rather than view autonomy in relationship to a void (individuals as *ex nihilo*), individuals emerge authoritative and powerful as part and parcel of an interconnected web of forces. Therefore, a crucial back-and-forth tug between the self and the various influences and authorities surrounding it is woven in the very fabric of what it means to be a fully attained and empowered individual.

This description, while not even derived from a cognitive science source, illustrates many of the key points of this paper. For example, in terms of agent terminology it states that the goal of a developing agent is to achieve authoritativeness (i.e., to internalize the role of authority) while at the same time conforming to certain types of authority stemming from the larger environment. Since we defined authority as the “processes or agents that create, maintain, and influence the conditions in which agents exist,” this description describes the outward development along the spiral in **Figure 1**. While all agents influence their living environment, it is the more authoritative—libertarian—agent that successfully can take a role as co-creator and co-maintainer of its environment. So co-creation—defined as working with the inherent dynamics of the world as opposed to frantically controlling and curtailing it—was an inherent part of early Chinese philosophy. In fact it corresponds to the Daoist key term “*Wu wei*,” which “means something like ‘act naturally,’ ‘effortless action,’ or ‘nonwillful action’” (Littlejohn, 2003). So the point of open-ended development is to learn “*Wu wei*” through a process of the internalization of authority insofar achievable given natural laws as highest authority.

The point to make here is not that early Chinese philosophy is an alternative to Western approaches to artificial cognitive systems research, but that our cultural biases limit our understanding. Accounting for these biases and learning from cultures without these particular (and probably other) biases

can help to inform the formulation of fundamental research roadmaps such as for artificial cognitive systems. We propose that putting the strengths of the right hemisphere (as summarized in **Table 1**) center-stage is an essential step to take artificial cognitive systems research out of the closed domain solutions afforded by left hemispheric approaches (and caricatures) of cognitive systems.

If the artificial cognitive systems community indeed tries to rid itself from its limiting biases and adopts approaches that puts the strengths of the right hemisphere and the open-ended development loop central, we have a suggestion for a suitable environment for artificial cognitive system development. This environment offers at the same time (1) many different agents and processes to relate with and care for, (2) many problems to solve and protocols to follow, and (3) an endless and unstoppable variety of novelty and change. This environment might have been an essential progenitor of our cultures because it approximates an ideal balance of reward signals to drive open-ended learning. So a robot that acts responsibly in this environment should be able to acquire the competences and moral development required to function responsibly in the rest of our societies. For that reason we suggest that robot labs should collaborate with ...low-tech self-sustaining farms where human, animals, vegetables, fruits, and grains flourish in one of the finest examples of what co-creation can offer.

REFERENCES

- Andringa, T. C. (2013). “Soundscape and its relation to core affect, appraisal, and motivation,” in *Presented at the AIA-DAGA 2013*, (Merano), 1511–1513.
- Andringa, T. C., and Langer, J. J. L. (2013). How pleasant sounds promote and annoying sounds impede health: a cognitive approach. *Int. J. Environ. Res. Public Health* 10, 1439–1461. doi: 10.3390/ijerph10041439
- Ardelt, M. (2000). Intellectual versus wisdom-related knowledge: the case for a different kind of learning in the later years of life. *Educ. Gerontol.* 26, 771–789. doi: 10.1080/036012700300001421
- Arnold, F. (1910). *Attention and Interest*. New York, NY: The Macmillan company.
- Axelsson, O., Nilsson, M. E., and Berglund, B. (2010). A principal components model of soundscape perception. *J. Acoust. Soc. Am.* 128, 2836–2846. doi: 10.1121/1.3493436
- Baldassarre, G. (2011). “What are intrinsic motivations? A biological perspective,” in *IEEE International Conference on Development and Learning*, Vol. 2, (Frankfurt), 1–8. doi: 10.1109/DEVLRN.2011.6037367
- Billington, J. H. (1980). *Fire in the Minds of Men*. New York, NY: Basic Books.
- Brindley, E. (2010). *Individualism in Early China*. Search.Ebscohost.com. Honolulu: University of Hawai'i Press. Available online at: <http://search.ebscohost.com/login.aspx?direct=true&scope=siteanddb=nlebk&anddb=nlabkandAN=336258>
- Brindley, E. (2011). “Individualism in classical Chinese thought,” in *Internet Encyclopedia of Philosophy*. Available online at: <http://www.iep.utm.edu/ind-chin/>
- Capra, F. (1997). *The Web of Life: a New Synthesis of Mind and Matter*. London: Flamingo.
- Chemero, A. (2003). An outline of a theory of affordances. *Ecol. Psychol.* 15, 181–195. doi: 10.1207/S15326969ECO1502_5
- Colby, A., Kohlberg, L., Gibbs, J., Lieberman, M., Fischer, K., and Saltzstein, H. D. (1983). A longitudinal study of moral judgment. *Monogr. Soc. Res. Child Dev.* 48, 1–124. doi: 10.2307/1165935
- Deci, E. L., and Ryan, R. M. (1987). The support of autonomy and the control of behavior. *J. Pers. Soc. Psychol.* 53, 1024–1037. doi: 10.1037/0022-3514.53.6.1024
- Fredrickson, B. L. (1998). What good are positive emotions? *Rev. Gen. Psychol.* 2, 300–319. doi: 10.1037/1089-2680.2.3.300
- Fredrickson, B. L., and Branigan, C. (2005). Positive emotions. *Cogn. Dev.* 15, 309–328. doi: 10.1080/108885-2014(00)00030-7
- Kuhn, D., Cheney, R., and Weinstock, M. (2000). The development of epistemological understanding. *Cogn. Dev.* 15, 309–328. doi: 10.1080/108885-2014(00)00030-7
- Kuppens, P., Champagne, D., and Tuerlinckx, F. (2012). The broaden the scope of attention and thought-action repertoires. *Cogn. Emot.* 19, 313–332. doi: 10.1080/02699930441000238
- Gomila, A., and Müller, V. C. (2012). Challenges for artificial cognitive systems. *J. Cogn. Sci.* 13, 453–470.
- Horney, K. (1945). *Our Inner Conflicts, a Constructive Theory of Neurosis*. New York, NY: W.W. Norton and Company Inc.
- Kaplan, S. (1995). The restorative benefits of nature: toward an integrative framework. *J. Environ. Psychol.* 15, 169–182. doi: 10.1016/0272-4944(95)90001-2
- Kauffman, S. (1995). *At Home in the Universe*. New York, NY: Oxford University Press.
- Kohlberg, L. (1971). Stages of moral development. *Moral Educ.* 23–92.
- Kohlberg, L. (1975). The cognitive-developmental approach to moral education. *Phi Delta Kappan* 56, 670–677.
- Kruglanski, A. W., and Webster, D. M. (1996). Motivated closing of the mind: “Seizing” and “Freezing”. *Psychol. Rev.* 103, 263–283. doi: 10.1037/0033-295X.103.2.263
- Maslow, A. H. (1943). A theory of human motivation. *Psychol. Rev.* 50, 370–396. doi: 10.1037/h0054346
- Maslow, A. H. (1954). “Motivation and Personality”, in *Chapter 11 of Self-actualizing People: A Study of Psychological Health*, ed. C. McReynolds (New York, NY: Harper and Row), 125–149.
- Maslow, A. H. (1962). *Toward a Psychology of Being*. New York, NY: D. van Nostrand company inc.
- Littlejohn, R. (2003). “Daoist philosophy.” *Internet Encyclopedia of Philosophy*. Available online at: <http://www.iep.utm.edu/daoism/>
- Maier, S. F., and Seligman, M. E. (1976). Learned helplessness: theory and evidence. *J. Exp. Psychol. Gen.* 105, 3–46. doi: 10.1037/0096-3445.105.1.3
- Malhotra, Y., Galletta, D. F., and Kirsch, L. J. (2008). How endogenous motivations influence user intentions: beyond the dichotomy of extrinsic and intrinsic user motivations. *J. Manag. Inf. Syst.* 25, 267–300. doi: 10.2753/MIS0742-1222250110
- Martin, M., Sadlo, G., and Stew, G. (2006). The phenomenon of boredom. *Qual. Res. Psychol.* 3, 193–211. doi: 10.1191/1478088706qrp066oa
- Maslow, A. H. (1943). A theory of human motivation. *Psychol. Rev.* 50, 370–396. doi: 10.1037/h0054346
- Malhotra, Y., Galletta, D. F., and Kirsch, L. J. (2008). How endogenous motivations influence user intentions: beyond the dichotomy of extrinsic and intrinsic user motivations. *J. Manag. Inf. Syst.* 25, 267–300. doi: 10.2753/MIS0742-1222250110
- Martin, M., Sadlo, G., and Stew, G. (2006). The phenomenon of boredom. *Qual. Res. Psychol.* 3, 193–211. doi: 10.1191/1478088706qrp066oa
- Maslow, A. H. (1943). A theory of human motivation. *Psychol. Rev.* 50, 370–396. doi: 10.1037/h0054346
- Malhotra, Y., Galletta, D. F., and Kirsch, L. J. (2008). How endogenous motivations influence user intentions: beyond the dichotomy of extrinsic and intrinsic user motivations. *J. Manag. Inf. Syst.* 25, 267–300. doi: 10.2753/MIS0742-1222250110
- Martin, M., Sadlo, G., and Stew, G. (2006). The phenomenon of boredom. *Qual. Res. Psychol.* 3, 193–211. doi: 10.1191/1478088706qrp066oa
- Maslow, A. H. (1943). A theory of human motivation. *Psychol. Rev.* 50, 370–396. doi: 10.1037/h0054346
- Malhotra, Y., Galletta, D. F., and Kirsch, L. J. (2008). How endogenous motivations influence user intentions: beyond the dichotomy of extrinsic and intrinsic user motivations. *J. Manag. Inf. Syst.* 25, 267–300. doi: 10.2753/MIS0742-1222250110
- Martin, M., Sadlo, G., and Stew, G. (2006). The phenomenon of boredom. *Qual. Res. Psychol.* 3, 193–211. doi: 10.1191/1478088706qrp066oa
- Maslow, A. H. (1943). A theory of human motivation. *Psychol. Rev.* 50, 370–396. doi: 10.1037/h0054346
- Malhotra, Y., Galletta, D. F., and Kirsch, L. J. (2008). How endogenous motivations influence user intentions: beyond the dichotomy of extrinsic and intrinsic user motivations. *J. Manag. Inf. Syst.* 25, 267–300. doi: 10.2753/MIS0742-1222250110
- Martin, M., Sadlo, G., and Stew, G. (2006). The phenomenon of boredom. *Qual. Res. Psychol.* 3, 193–211. doi: 10.1191/1478088706qrp066oa
- Maslow, A. H. (1943). A theory of human motivation. *Psychol. Rev.* 50, 370–396. doi: 10.1037/h0054346
- Malhotra, Y., Galletta, D. F., and Kirsch, L. J. (2008). How endogenous motivations influence user intentions: beyond the dichotomy of extrinsic and intrinsic user motivations. *J. Manag. Inf. Syst.* 25, 267–300. doi: 10.2753/MIS0742-1222250110
- Martin, M., Sadlo, G., and Stew, G. (2006). The phenomenon of boredom. *Qual. Res. Psychol.* 3, 193–211. doi: 10.1191/1478088706qrp066oa
- Maslow, A. H. (1943). A theory of human motivation. *Psychol. Rev.* 50, 370–396. doi: 10.1037/h0054346
- Malhotra, Y., Galletta, D. F., and Kirsch, L. J. (2008). How endogenous motivations influence user intentions: beyond the dichotomy of extrinsic and intrinsic user motivations. *J. Manag. Inf. Syst.* 25, 267–300. doi: 10.2753/MIS0742-1222250110
- Martin, M., Sadlo, G., and Stew, G. (2006). The phenomenon of boredom. *Qual. Res. Psychol.* 3, 193–211. doi: 10.1191/1478088706qrp066oa
- Maslow, A. H. (1943). A theory of human motivation. *Psychol. Rev.* 50, 370–396. doi: 10.1037/h0054346
- Malhotra, Y., Galletta, D. F., and Kirsch, L. J. (2008). How endogenous motivations influence user intentions: beyond the dichotomy of extrinsic and intrinsic user motivations. *J. Manag. Inf. Syst.* 25, 267–300. doi: 10.2753/MIS0742-1222250110
- Martin, M., Sadlo, G., and Stew, G. (2006). The phenomenon of boredom. *Qual. Res. Psychol.* 3, 193–211. doi: 10.1191/1478088706qrp066oa
- Maslow, A. H. (1943). A theory of human motivation. *Psychol. Rev.* 50, 370–396. doi: 10.1037/h0054346
- Malhotra, Y., Galletta, D. F., and Kirsch, L. J. (2008). How endogenous motivations influence user intentions: beyond the dichotomy of extrinsic and intrinsic user motivations. *J. Manag. Inf. Syst.* 25, 267–300. doi: 10.2753/MIS0742-1222250110
- Martin, M., Sadlo, G., and Stew, G. (2006). The phenomenon of boredom. *Qual. Res. Psychol.* 3, 193–211. doi: 10.1191/1478088706qrp066oa
- Maslow, A. H. (1943). A theory of human motivation. *Psychol. Rev.* 50, 370–396. doi: 10.1037/h0054346
- Malhotra, Y., Galletta, D. F., and Kirsch, L. J. (2008). How endogenous motivations influence user intentions: beyond the dichotomy of extrinsic and intrinsic user motivations. *J. Manag. Inf. Syst.* 25, 267–300. doi: 10.2753/MIS0742-1222250110
- Martin, M., Sadlo, G., and Stew, G. (2006). The phenomenon of boredom. *Qual. Res. Psychol.* 3, 193–211. doi: 10.1191/1478088706qrp066oa
- Maslow, A. H. (1943). A theory of human motivation. *Psychol. Rev.* 50, 370–396. doi: 10.1037/h0054346
- Malhotra, Y., Galletta, D. F., and Kirsch, L. J. (2008). How endogenous motivations influence user intentions: beyond the dichotomy of extrinsic and intrinsic user motivations. *J. Manag. Inf. Syst.* 25, 267–300. doi: 10.2753/MIS0742-1222250110
- Martin, M., Sadlo, G., and Stew, G. (2006). The phenomenon of boredom. *Qual. Res. Psychol.* 3, 193–211. doi: 10.1191/1478088706qrp066oa
- Maslow, A. H. (1943). A theory of human motivation. *Psychol. Rev.* 50, 370–396. doi: 10.1037/h0054346
- Malhotra, Y., Galletta, D. F., and Kirsch, L. J. (2008). How endogenous motivations influence user intentions: beyond the dichotomy of extrinsic and intrinsic user motivations. *J. Manag. Inf. Syst.* 25, 267–300. doi: 10.2753/MIS0742-1222250110
- Martin, M., Sadlo, G., and Stew, G. (2006). The phenomenon of boredom. *Qual. Res. Psychol.* 3, 193–211. doi: 10.1191/1478088706qrp066oa
- Maslow, A. H. (1943). A theory of human motivation. *Psychol. Rev.* 50, 370–396. doi: 10.1037/h0054346
- Malhotra, Y., Galletta, D. F., and Kirsch, L. J. (2008). How endogenous motivations influence user intentions: beyond the dichotomy of extrinsic and intrinsic user motivations. *J. Manag. Inf. Syst.* 25, 267–300. doi: 10.2753/MIS0742-1222250110
- Martin, M., Sadlo, G., and Stew, G. (2006). The phenomenon of boredom. *Qual. Res. Psychol.* 3, 193–211. doi: 10.1191/1478088706qrp066oa
- Maslow, A. H. (1943). A theory of human motivation. *Psychol. Rev.* 50, 370–396. doi: 10.1037/h0054346
- Malhotra, Y., Galletta, D. F., and Kirsch, L. J. (2008). How endogenous motivations influence user intentions: beyond the dichotomy of extrinsic and intrinsic user motivations. *J. Manag. Inf. Syst.* 25, 267–300. doi: 10.2753/MIS0742-1222250110
- Martin, M., Sadlo, G., and Stew, G. (2006). The phenomenon of boredom. *Qual. Res. Psychol.* 3, 193–211. doi: 10.1191/1478088706qrp066oa
- Maslow, A. H. (1943). A theory of human motivation. *Psychol. Rev.* 50, 370–396. doi: 10.1037/h0054346
- Malhotra, Y., Galletta, D. F., and Kirsch, L. J. (2008). How endogenous motivations influence user intentions: beyond the dichotomy of extrinsic and intrinsic user motivations. *J. Manag. Inf. Syst.* 25, 267–300. doi: 10.2753/MIS0742-1222250110
- Martin, M., Sadlo, G., and Stew, G. (2006). The phenomenon of boredom. *Qual. Res. Psychol.* 3, 193–211. doi: 10.1191/1478088706qrp066oa
- Maslow, A. H. (1943). A theory of human motivation. *Psychol. Rev.* 50, 370–396. doi: 10.1037/h0054346
- Malhotra, Y., Galletta, D. F., and Kirsch, L. J. (2008). How endogenous motivations influence user intentions: beyond the dichotomy of extrinsic and intrinsic user motivations. *J. Manag. Inf. Syst.* 25, 267–300. doi: 10.2753/MIS0742-1222250110
- Martin, M., Sadlo, G., and Stew, G. (2006). The phenomenon of boredom. *Qual. Res. Psychol.* 3, 193–211. doi: 10.1191/1478088706qrp066oa
- Maslow, A. H. (1943). A theory of human motivation. *Psychol. Rev.* 50, 370–396. doi: 10.1037/h0054346
- Malhotra, Y., Galletta, D. F., and Kirsch, L. J. (2008). How endogenous motivations influence user intentions: beyond the dichotomy of extrinsic and intrinsic user motivations. *J. Manag. Inf. Syst.* 25, 267–300. doi: 10.2753/MIS0742-1222250110
- Martin, M., Sadlo, G., and Stew, G. (2006). The phenomenon of boredom. *Qual. Res. Psychol.* 3, 193–211. doi: 10.1191/1478088706qrp066oa
- Maslow, A. H. (1943). A theory of human motivation. *Psychol. Rev.* 50, 370–396. doi: 10.1037/h0054346
- Malhotra, Y., Galletta, D. F., and Kirsch, L. J. (2008). How endogenous motivations influence user intentions: beyond the dichotomy of extrinsic and intrinsic user motivations. *J. Manag. Inf. Syst.* 25, 267–300. doi: 10.2753/MIS0742-1222250110
- Martin, M., Sadlo, G., and Stew, G. (2006). The phenomenon of boredom. *Qual. Res. Psychol.* 3, 193–211. doi: 10.1191/1478088706qrp066oa
- Maslow, A. H. (1943). A theory of human motivation. *Psychol. Rev.* 50, 370–396. doi: 10.1037/h0054346
- Malhotra, Y., Galletta, D. F., and Kirsch, L. J. (2008). How endogenous motivations influence user intentions: beyond the dichotomy of extrinsic and intrinsic user motivations. *J. Manag. Inf. Syst.* 25, 267–300. doi: 10.2753/MIS0742-1222250110
- Martin, M., Sadlo, G., and Stew, G. (2006). The phenomenon of boredom. *Qual. Res. Psychol.* 3, 193–211. doi: 10.1191/1478088706qrp066oa
- Maslow, A. H. (1943). A theory of human motivation. *Psychol. Rev.* 50, 370–396. doi: 10.1037/h0054346
- Malhotra, Y., Galletta, D. F., and Kirsch, L. J. (2008). How endogenous motivations influence user intentions: beyond the dichotomy of extrinsic and intrinsic user motivations. *J. Manag. Inf. Syst.* 25, 267–300. doi: 10.2753/MIS0742-1222250110
- Martin, M., Sadlo, G., and Stew, G. (2006). The phenomenon of boredom. *Qual. Res. Psychol.* 3, 193–211. doi: 10.1191/1478088706qrp066oa
- Maslow, A. H. (1943). A theory of human motivation. *Psychol. Rev.* 50, 370–396. doi: 10.1037/h0054346
- Malhotra, Y., Galletta, D. F., and Kirsch, L. J. (2008). How endogenous motivations influence user intentions: beyond the dichotomy of extrinsic and intrinsic user motivations. *J. Manag. Inf. Syst.* 25, 267–300. doi: 10.2753/MIS0742-1222250110
- Martin, M., Sadlo, G., and Stew, G. (2006). The phenomenon of boredom. *Qual. Res. Psychol.* 3, 193–211. doi: 10.1191/1478088706qrp066oa
- Maslow, A. H. (1943). A theory of human motivation. *Psychol. Rev.* 50, 370–396. doi: 10.1037/h0054346
- Malhotra, Y., Galletta, D. F., and Kirsch, L. J. (2008). How endogenous motivations influence user intentions: beyond the dichotomy of extrinsic and intrinsic user motivations. *J. Manag. Inf. Syst.* 25, 267–300. doi: 10.2753/MIS0742-1222250110
- Martin, M., Sadlo, G., and Stew, G. (2006). The phenomenon of boredom. *Qual. Res. Psychol.* 3, 193–211. doi: 10.1191/1478088706qrp066oa
- Maslow, A. H. (1943). A theory of human motivation. *Psychol. Rev.* 50, 370–396. doi: 10.1037/h0054346
- Malhotra, Y., Galletta, D. F., and Kirsch, L. J. (2008). How endogenous motivations influence user intentions: beyond the dichotomy of extrinsic and intrinsic user motivations. *J. Manag. Inf. Syst.* 25, 267–300. doi: 10.2753/MIS0742-1222250110
- Martin, M., Sadlo, G., and Stew, G. (2006). The phenomenon of boredom. *Qual. Res. Psychol.* 3, 193–211. doi: 10.1191/1478088706qrp066oa
- Maslow, A. H. (1943). A theory of human motivation. *Psychol. Rev.* 50, 370–396. doi: 10.1037/h0054346
- Malhotra, Y., Galletta, D. F., and Kirsch, L. J. (2008). How endogenous motivations influence user intentions: beyond the dichotomy of extrinsic and intrinsic user motivations. *J. Manag. Inf. Syst.* 25, 267–300. doi: 10.2753/MIS0742-1222250110
- Martin, M., Sadlo, G., and Stew, G. (2006). The phenomenon of boredom. *Qual. Res. Psychol.* 3, 193–211. doi: 10.1191/1478088706qrp066oa
- Maslow, A. H. (1943). A theory of human motivation. *Psychol. Rev.* 50, 370–396. doi: 10.1037/h0054346
- Malhotra, Y., Galletta, D. F., and Kirsch, L. J. (2008). How endogenous motivations influence user intentions: beyond the dichotomy of extrinsic and intrinsic user motivations. *J. Manag. Inf. Syst.* 25, 267–300. doi: 10.2753/MIS0742-1222250110
- Martin, M., Sadlo, G., and Stew, G. (2006). The phenomenon of boredom. *Qual. Res. Psychol.* 3, 193–211. doi: 10.1191/1478088706qrp066oa
- Maslow, A. H. (1943). A theory of human motivation. *Psychol. Rev.* 50, 370–396. doi: 10.1037/h0054346
- Malhotra, Y., Galletta, D. F., and Kirsch, L. J. (2008). How endogenous motivations influence user intentions: beyond the dichotomy of extrinsic and intrinsic user motivations. *J. Manag. Inf. Syst.* 25, 267–300. doi: 10.2753/MIS0742-1222250110
- Martin, M., Sadlo, G., and Stew, G. (2006). The phenomenon of boredom. *Qual. Res. Psychol.* 3, 193–211. doi: 10.1191/1478088706qrp066oa
- Maslow, A. H. (1943). A theory of human motivation. *Psychol. Rev.* 50, 370–396. doi: 10.1037/h0054346
- Malhotra, Y., Galletta, D. F., and Kirsch, L. J. (2008). How endogenous motivations influence user intentions: beyond the dichotomy of extrinsic and intrinsic user motivations. *J. Manag. Inf. Syst.* 25, 267–300. doi: 10.2753/MIS0742-1222250110
- Martin, M., Sadlo, G., and Stew, G. (2006). The phenomenon of boredom. *Qual. Res. Psychol.* 3, 193–211. doi: 10.1191/1478088706qrp066oa
- Maslow, A. H. (1943). A theory of human motivation. *Psychol. Rev.* 50, 370–396. doi: 10.1037/h0054346
- Malhotra, Y., Galletta, D. F., and Kirsch, L. J. (2008). How endogenous motivations influence user intentions: beyond the dichotomy of extrinsic and intrinsic user motivations. *J. Manag. Inf. Syst.* 25, 267–300. doi: 10.2753/MIS0742-1222250110
- Martin, M., Sadlo, G., and Stew, G. (2006). The phenomenon of boredom. *Qual. Res. Psychol.* 3, 193–211. doi: 10.1191/1478088706qrp066oa
- Maslow, A. H. (1943). A theory of human motivation. *Psychol. Rev.* 50, 370–396. doi: 10.1037/h0054346
- Malhotra, Y., Galletta, D. F., and Kirsch, L. J. (2008). How endogenous motivations influence user intentions: beyond the dichotomy of extrinsic and intrinsic user motivations. *J. Manag. Inf. Syst.* 25, 267–300. doi: 10.2753/MIS0742-1222250110
- Martin, M., Sadlo, G., and Stew, G. (2006). The phenomenon of boredom. *Qual. Res. Psychol.* 3, 193–211. doi: 10.1191/1478088706qrp066oa
- Maslow, A. H. (1943). A theory of human motivation. *Psychol. Rev.* 50, 370–396. doi: 10.1037/h0054346
- Malhotra, Y., Galletta, D. F., and Kirsch, L. J. (2008). How endogenous motivations influence user intentions: beyond the dichotomy of extrinsic and intrinsic user motivations. *J. Manag. Inf. Syst.* 25, 267–300. doi: 10.2753/MIS0742-1222250110
- Martin, M., Sadlo, G., and Stew, G. (2006). The phenomenon of boredom. *Qual. Res. Psychol.* 3, 193–211. doi: 10.1191/1478088706qrp066oa
- Maslow, A. H. (1943). A theory of human motivation. *Psychol. Rev.* 50, 370–396. doi: 10.1037/h0054346
- Malhotra, Y., Galletta, D. F., and Kirsch, L. J. (2008). How endogenous motivations influence user intentions: beyond the dichotomy of extrinsic and intrinsic user motivations. *J. Manag. Inf. Syst.* 25, 267–300. doi: 10.2753/MIS0742-1222250110
- Martin, M., Sadlo, G., and Stew, G. (2006). The phenomenon of boredom. *Qual. Res. Psychol.* 3, 193–211. doi: 10.1191/1478088706qrp066oa
- Maslow, A. H. (1943). A theory of human motivation. *Psychol. Rev.* 50, 370–396. doi: 10.1037/h0054346
- Malhotra, Y., Galletta, D. F., and Kirsch, L. J. (2008). How endogenous motivations influence user intentions: beyond the dichotomy of extrinsic and intrinsic user motivations. *J. Manag. Inf. Syst.* 25, 267–300. doi: 10.2753/MIS0742-1222250110
- Martin, M., Sadlo, G., and Stew, G. (2006). The phenomenon of boredom. *Qual. Res. Psychol.* 3, 193–211. doi: 10.1191/1478088706qrp066oa
- Maslow, A. H. (1943). A theory of human motivation. *Psychol. Rev.* 50, 370–396. doi: 10.1037/h0054346
- Malhotra, Y., Galletta, D. F., and Kirsch, L. J. (2008). How endogenous motivations influence user intentions: beyond the dichotomy of extrinsic and intrinsic user motivations. *J. Manag. Inf. Syst.* 25, 267–300. doi: 10.2753/MIS0742-1222250110
- Martin, M., Sadlo, G., and Stew, G. (2006). The phenomenon of boredom. *Qual. Res. Psychol.* 3, 193–211. doi: 10.1191/1478088706qrp066oa
- Maslow, A. H. (1943). A theory of human motivation. *Psychol. Rev.* 50, 370–396. doi: 10.1037/h0054346
- Malhotra, Y., Galletta, D. F., and Kirsch, L. J. (2008). How endogenous motivations influence user intentions: beyond the dichotomy of extrinsic and intrinsic user motivations. *J. Manag. Inf. Syst.* 25, 267–300. doi: 10.2753/MIS0742-1222250110
- Martin, M., Sadlo, G., and Stew, G. (2006). The phenomenon of boredom. *Qual. Res. Psychol.* 3, 193–211. doi: 10.1191/1478088706qrp066oa
- Maslow,

- McCrae, R. R., and Sutin, A. R. (2009). "Openness to experience," in *Handbook of Individual Differences in Social Behavior*, eds M. R. Leary and R. H. Hoyle (New York, NY: Guilford), 257–273. doi: 10.1007/s00018-010-0311-0
- McGilchrist, I. (2010). *The Master and His Emissary: the Divided Brain and the Making of the Western World*. New Haven, CT: Yale University Press. Available online at: http://www.iainmcgilchrist.com/TMAHE/biblio/Bibliography_The_Master_and_his_Emissary.pdf
- Mora, T., and Bialek, W. (2011). Are biological systems poised at criticality? *J. Stat. Phys.* 144, 268–302. doi: 10.1007/s10955-011-0229-4
- Nakamura, J., and Csikszentmihalyi, M. (2002). "The concept of flow," in *Handbook of positive psychology*, eds C. R. Snyder and S. J. Lopez (New York, NY: Oxford University Press), 89–105.
- Perkins, R. E., and Hill, A. B. (1985). Cognitive and affective aspects of boredom. *Br. J. Psychol.* 76, 221–234. doi: 10.1111/j.2044-8295.1985.tb01946.x
- Richardson, M. J., and Pasupathi, M. (2005). "Young and growing wiser: wisdom during adolescence and young adulthood," in *A Handbook of Wisdom: Psychological Perspectives*, eds R. J. Sternberg and J. Jordan (Cambridge; New York, NY: Cambridge University Press), 139–159. doi: 10.1017/CBO9780511610486.007
- Rowson, J., and McGilchrist, I. (2013). "Divided brain, divided world." *Action and Research Center (RSA)*. Available online at: www.thersa.org.
- Russell, B. (1938). *Power*. 1st Edn. London: George Allen and Unwin.
- Russell, J. (2003). Core affect and the psychological construction of emotion. *Psychol. Rev.* 110, 145–172. doi: 10.1037/0033-295X.110.1.145
- Russell, J. A., Ward, L. M., and Pratt, G. (1981). Affective quality attributed to environments: a factor analytic study. *Environ. Behav.* 13, 259–288. doi: 10.1177/0013916581133001
- Ryan, R. M., and Connell, J. P. (1989). Perceived locus of causality and internalization: examining reasons for acting in two domains. *J. Pers. Psychol.* 57, 749–761. doi: 10.1037/0022-3514.57.5.749
- Ryan, R. M., and Deci, E. L. (2000). Intrinsic and extrinsic motivations: classic definitions and new directions. *Contemp. Educ. Psychol.* 25, 54–67. doi: 10.1006/ceps.1999.1020
- Silvia, P. J. (2008). Interest—the curious emotion. *Curr. Dir. Psychol. Sci.* 17, 57–60. doi: 10.1111/j.1467-8721.2008.00548.x
- Staudinger, U. M., and Glück, J. (2011). Psychological wisdom research: commonalities and differences in a growing field. *Annu. Rev. Psychol.* 62, 215–241. doi: 10.1146/annurev.psych.121208.131659
- Stenner, K. (2005). *The Authoritarian Dynamic*. 1st Edn. New York, NY: Cambridge University Press. doi: 10.1017/CBO9780511614712
- Stenner, K. (2009). Conservatism, context-dependence, and cognitive incapacity. *Psychol. Inq.* 20, 189–195. doi: 10.1080/10478400903123994
- Sternberg, R. J. (1998). A balance theory of wisdom. *Rev. Gen. Psychol.* 2, 347–365. doi: 10.1037/1089-2680.2.4.347
- van Rossum, E. J., and Hamer, R. N. (2010). "Students' conceptions of learning," in *Chapter 1 of The Meaning of Learning and Knowing*, ed J. Vermunt (Utrecht: University of Utrecht).
- Vygotsky, L. L. S. (1978). *Mind in Society*. ed M. Cole, V. John-Steiner, S. Scribner, and E. Souberman. London: Harvard University Press.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 17 May 2013; accepted: 30 September 2013; published online: 22 October 2013.

Citation: Andringa TC, van den Bosch KA and Vlaskamp C (2013) Learning autonomy in two or three steps: linking open-ended development, authority, and agency to motivation. *Front. Psychol.* 4:766. doi: 10.3389/fpsyg.2013.00766
This article was submitted to Cognitive Science, a section of the journal *Frontiers in Psychology*.

Copyright © 2013 Andringa, van den Bosch and Vlaskamp. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Emergent structured transition from variation to repetition in a biologically-plausible model of learning in basal ganglia

Ashvin Shah * and Kevin N. Gurney

Department of Psychology, University of Sheffield, Sheffield, UK

Edited by:

Gianluca Baldassarre, Italian National Research Council, Italy

Reviewed by:

Fred H. Hamker, Chemnitz University of Technology, Germany
Vincenzo G. Fiore, UCL Institute of Neurology, UK
Valerio Sperati, Consiglio Nazionale delle Ricerche, Italy

***Correspondence:**

Ashvin Shah, Department of Psychology, University of Sheffield, Western Bank, Sheffield S10 2TP, UK
e-mail: a.shah@sheffield.ac.uk

Often, when animals encounter an unexpected sensory event, they transition from executing a variety of movements to repeating the movement(s) that may have caused the event. According to a recent theory of action discovery (Redgrave and Gurney, 2006), repetition allows the animal to represent those movements, and the outcome, as an action for later recruitment. The transition from variation to repetition often follows a non-random, structured, pattern. While the structure of the pattern can be explained by sophisticated cognitive mechanisms, simpler mechanisms based on dopaminergic modulation of basal ganglia (BG) activity are thought to underlie action discovery (Redgrave and Gurney, 2006). In this paper we ask the question: can simple BG-mediated mechanisms account for a structured transition from variation to repetition, or are more sophisticated cognitive mechanisms always necessary? To address this question, we present a computational model of BG-mediated biasing of behavior. In our model, unlike most other models of BG function, the BG biases behavior through modulation of cortical response to excitation; many possible movements are represented by the cortical area; and excitation to the cortical area is topographically-organized. We subject the model to simple reaching tasks, inspired by behavioral studies, in which a location to which to reach must be selected. Locations within a target area elicit a reinforcement signal. A structured transition from variation to repetition emerges from simple BG-mediated biasing of cortical response to excitation. We show how the structured pattern influences behavior in simple and complicated tasks. We also present analyses that describe the structured transition from variation to repetition due to BG-mediated biasing and from biasing that would be expected from a type of cognitive biasing, allowing us to compare behavior resulting from these types of biasing and make connections with future behavioral experiments.

Keywords: action discovery, reinforcement, basal ganglia, variation, repetition

1. INTRODUCTION

Animals are capable of executing a huge variety of movements but, importantly, they can discover the specific movements that affect the environment in predictable ways and represent them as *actions* for later recruitment. Redgrave, Gurney, and colleagues have suggested that this occurs through a process they refer to as *action discovery* (Redgrave and Gurney, 2006; Redgrave et al., 2008, 2011, 2013; Gurney et al., 2013). Action discovery begins when an animal is executing movements within some context and an unexpected salient sensory event (such as a light flash) occurs. The unexpected sensory event causes a short-latency phasic increase in dopamine (DA) neuron activity (henceforth referred to simply as *DA activity*). Through its influence on the basal ganglia (BG)—a group of interconnected subcortical structures which, in turn, influence cortical activity—the increase in DA activity can help bias the animal to repeat the movements that preceded the unexpected sensory event under the same contextual circumstances. This *repetition bias* (Redgrave and Gurney, 2006) allows associative networks in the brain to learn and encode the

movements as an action because it causes a frequent and reliable presentation of context, movements, and the sensory event as the outcome of those movements.

This transition from executing a variety of movements to repeating just one or a subset of movements often follows a non-random, structured, pattern. For example, consider a spatial task such that reaching to a specific location results in the outcome. Here, one type of structured transition from variation to repetition occurs if the animal gradually refines its movements so that movements that are further from the location decrease in frequency earlier than movements that are closer to the location.

The non-random structure of the transition from variation to repetition can be explained with “intelligent” or sophisticated cognitive mechanisms, e.g., by using an estimation of the range of movements that cause the outcome that gets more and more precise with repeated occurrences of the outcome. Similarly, other types of a structured transition may rely on other sophisticated notions such as optimality or uncertainty (e.g., Dearden et al. 1998; Dimitrakakis 2006; Simsek and Barto 2006). However, the

process of action discovery is thought to be mediated primarily by simpler mechanisms involving DA modulation of the BG, and not sophisticated cognitive mechanisms. In this paper we ask the question, can *simple* BG-mediated mechanisms guide a structured transition from variation to repetition, or must sophisticated cognitive mechanisms always be recruited? To address this question, we present a computational model of BG-mediated biasing of behavior.

Our model will necessarily deal with a specific and, therefore, limited example of action discovery and so to establish its status, we now outline the model's wider context comprising various broad categories of action. For example, one type of action might involve making a particular gesture with the hand (as in sign language or hand signaling), regardless of the precise spatial location of the hand, and no environmental object is targeted. Another type of action involves manipulating objects in the environment (such as flipping a light switch or typing out a password). In this instance, space is weakly implicit (the objects are located somewhere); the key feature is the target object identity and its manipulation. In this paper, we focus on an explicitly spatial task: the relatively simple action of moving an end-effector to a particular spatial location. In the model task, a movement end-point to which to move must be selected. End-points that correspond to a target location elicit a reinforcement signal, and, importantly, reinforcement is not contingent on movement trajectory. The model task is inspired by behavioral counterparts we have used to study action discovery in which participants manipulate a joystick to find an invisible target area in the workspace (Stafford et al., 2012, 2013; Thirkettle et al., 2013a,b). While there may be "gestural" aspects of action in the behavioral task, in the model we ignore these and focus only on the spatial location of movement end-point.

In the next few paragraphs, we describe features of neural processing which our model incorporates that many other models of the BG do not. Biological theories of BG function suggest that the BG bias behavior not through direct excitation of their efferent targets, but, rather, through the selective relaxation of inhibition (i.e., disinhibition) of their efferent targets (Chevalier and Deniau, 1990; Mink, 1996; Redgrave et al., 2011). When the BG are presented with multiple signals, each representing an action or movement, these signals will have different activity levels signifying the urgency or *salience* of the "action request." BG are supposed to process each signal through a neural population or *channel*, and inter-channel connections facilitate competitive processes resulting in suppression of BG output (inhibition) on high salience channels and increased output on the low salience channels (Gurney et al., 2001a,b; Humphries and Gurney, 2002; Prescott et al., 2006). Many models of BG function focus on how the multiple signals presented to the BG are transformed to the activity of the BG's output nucleus. Action selection in these models is then based on the latter's activity (e.g., Gurney et al. 2001a,b, 2004; Joel et al. 2002; Daw et al. 2005; Shah and Barto 2009). However, one important feature of our model is that it also takes into account the pattern of excitation from other areas to the BG's efferent targets (see also Humphries and Gurney 2002; Cohen and Frank 2009; Baldassarre et al. 2013). Thus, behavior results from BG modulation of their efferent target's response to excitation

patterns, and is not just a mirror of the activity of the BG's output nucleus.

Further, many models of BG function focus on how the BG select from a small number of abstract independent behaviors (e.g., Gurney et al. 2001b; Daw et al. 2005; Cohen and Frank 2009; Shah and Barto 2009). While such representations may be appropriate for some behavioral tasks in experimental psychology, in ethological action discovery, the space of activities from which to select may be larger and adhere to some inherent topology. In our model, candidate locations to which to move are represented by a large number of topographically-organized neurons in cortex so that neighboring spatial locations are represented by neighboring neurons. Excitation to cortex follows a pattern in which all neurons are weakly excited initially, and that pattern evolves so that eventually only one neuron is excited strongly. This pattern is inspired by neural activity observed in perceptual decision-making tasks (Britten et al., 1992; Platt and Glimcher, 1999; Huk and Shadlen, 2005; Gold and Shadlen, 2007), and as suggested by evidence accumulation models of decision-making (Bogacz et al., 2006; Lepora et al., 2012).

We hypothesize that because the BG bias behavior by modulating cortical response to excitation, and that that excitation follows a structured pattern, simple BG-mediated biasing can result in a structured transition from variation to repetition in action discovery. Sophisticated cognitive mechanisms are not necessarily required to develop a structured transition.

In addition, behavioral biasing in action discovery is not thought to be driven by "extrinsic motivations" that are based on rewarding consequences and that dictate reinforcement in many types of operant conditioning tasks (Thorndike, 1911; Skinner, 1938) and computational reinforcement learning (RL) (Bertsekas and Tsitsiklis, 1996; Sutton and Barto, 1998). Rather, "intrinsic motivations" (Oudeyer and Kaplan, 2007; Baldassarre, 2011; Barto, 2013; Barto et al., 2013; Gottlieb et al., 2013; Gurney et al., 2013) that are triggered by the occurrence of an unexpected sensory event may drive DA activity and thus behavioral biasing in action discovery (Redgrave and Gurney, 2006; Redgrave et al., 2008, 2011, 2013; Gurney et al., 2013; Mirolli et al., 2013). In such cases, if the outcome does not represent or predict an extrinsically-rewarding event, reinforcement decreases as associative networks in the brain learn to predict its occurrence (Redgrave and Gurney, 2006; Redgrave et al., 2011). Rather than implement a model of prediction explicitly, we approximate its effects with a simple model of habituation in which the rate of reinforcement decreases as the target location is repeatedly hit (Marsland, 2009). This habituation model approximates the dependence of DA activity on outcome predictability in action discovery (Redgrave and Gurney, 2006; Redgrave et al., 2011), and is similar to that used in neural network models of novelty detection (Marsland, 2009).

In this paper, we use computational models to demonstrate that simple BG-mediated mechanisms can bias behavior, via their modulation of cortical response to a pattern of excitation, such that the transition from variation to repetition follows a structured pattern. We describe this structured pattern and show how it, along with the effects of habituation, lead to behavioral patterns in tasks in which one target area delivers a reinforcement

signal, two target areas deliver reinforcement, or the target area that delivers reinforcement changes location. These experiments lead to predictions as to the type of behavior that would be expected when only simple BG-mediated mechanisms, and not more sophisticated cognitive mechanisms, bias behavior. We also run models that mimic a simple form of transition from variation to repetition that would be expected under sophisticated cognitive mechanisms by subsuming the effects of those mechanisms in a phenomenological way. In order to make contact with future behavioral experiments, we develop a novel characterization of behavioral trends which links these trends to underlying neural mechanisms that dictate different forms of biasing.

2. METHODS

We use a computational model, based on established models (Gurney et al., 2001a,b; Humphries and Gurney, 2002), to control movement selection in a task that simulates reaching or pointing to specific target spatial locations. We provide here a conceptual overview of its mechanics; detailed equations are provided in the Supplementary section.

The model is a neural network model with leaky-integrator neuron units (henceforth referred to as “neurons” for brevity), the activities of which represent conglomerate neural firing rate of a group of neurons (Gurney et al., 2001a,b). Each brain area in the model, except for the area labeled “Context,” consists of 196 neurons spatially arranged in a 14×14 grid. Each neuron in each area is part of an “action channel” (Gurney et al., 2001a,b; Humphries and Gurney, 2002) such that its location in the grid corresponds to a movement toward the corresponding location of a two-dimensional workspace. For the purposes of this model, the workspace is of dimensions 14×14 units. Most projections from one area to another are one-to-one and not plastic; exceptions will be explicitly noted.

Figure 1 illustrates the gross architecture of the model. In brief, the end-point location of a movement, X_M , is determined by the activities of neurons in “M (Cortex).” These neurons are excited by an exploratory mechanism, “E (Explorer),” and are engaged in positive feedback loops with neurons in “T (Thalamus).” The basal ganglia (BG, gray boxes) send inhibitory projections to Thalamus neurons, and they modulate the gain of the Cortex-Thalamus positive feedback loops (Chambers et al., 2011) through selective disinhibition of Thalamus neurons. Cortex and Thalamus represent grids of neurons that correspond to motor-related areas of cortex and thalamus, respectively.

2.1. EXCITATORY INPUTS TO THE NEURAL NETWORK

There are two sources of excitatory input to the neural network. The first is labeled “C (Context)” and represents the context, such as participating in the current experiment. There is only one context for the results reported in this paper. Thus, Context consists of a single neuron with an output activity set to a constant value. Context influences BG activity through one-to-all projections to areas D1, D2, and STN. Projections to D1 and D2 are plastic and represent a context-dependent biasing of movements, as described in the subsection “Biasing of behavior.”

The second source of excitatory input is “E (Explorer),” which provides excitation to Cortex which, in turn, is responsible for

movement. The Explorer is the source of variation required to explore the space of possible movements. This variation may be more or less random or structured according to the strategy used. However, these strategies are devised by other mechanisms, not explicitly modeled here, and we simply aim to capture the effects of such strategies in the Explorer.

In this paper, the Explorer is inspired by a range of experimental data. First, recordings in some areas of parietal cortices (Anderson and Buneo, 2002) show activation of neurons corresponding to a decision to make a movement that terminates at the location represented by those neurons. Further, several experimental studies, (Britten et al., 1992; Platt and Glimcher, 1999; Huk and Shadlen, 2005; Gold and Shadlen, 2007) show that neurons representing different decisions are weakly active early in the decision-making process. The activities of some neurons—corresponding to the executed decision in these experiments—increase at a greater rate than that of other neurons.

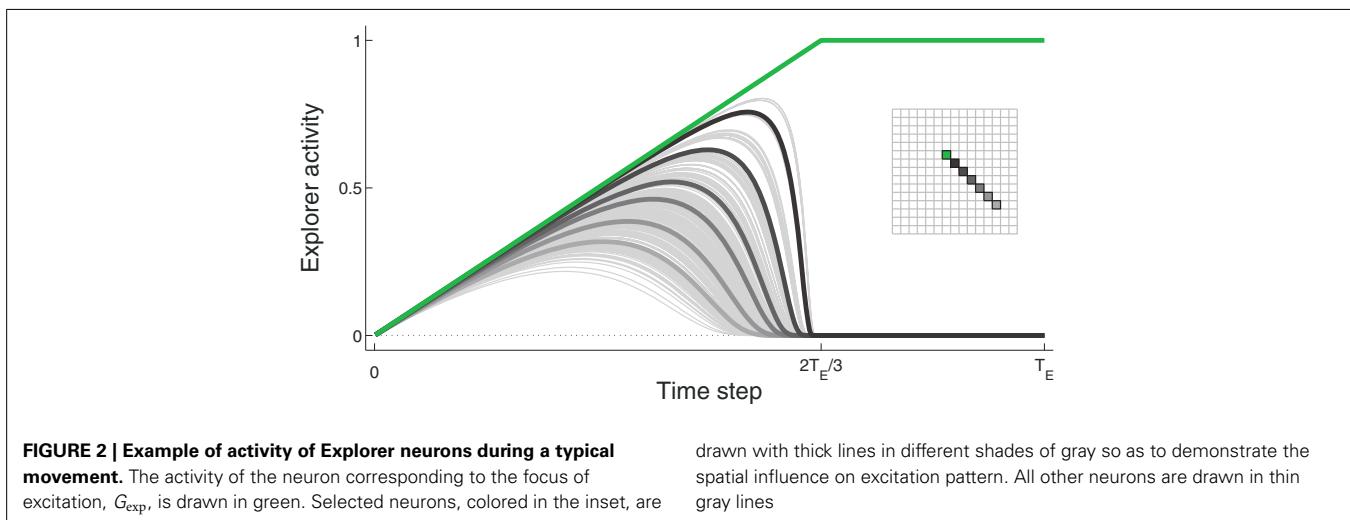
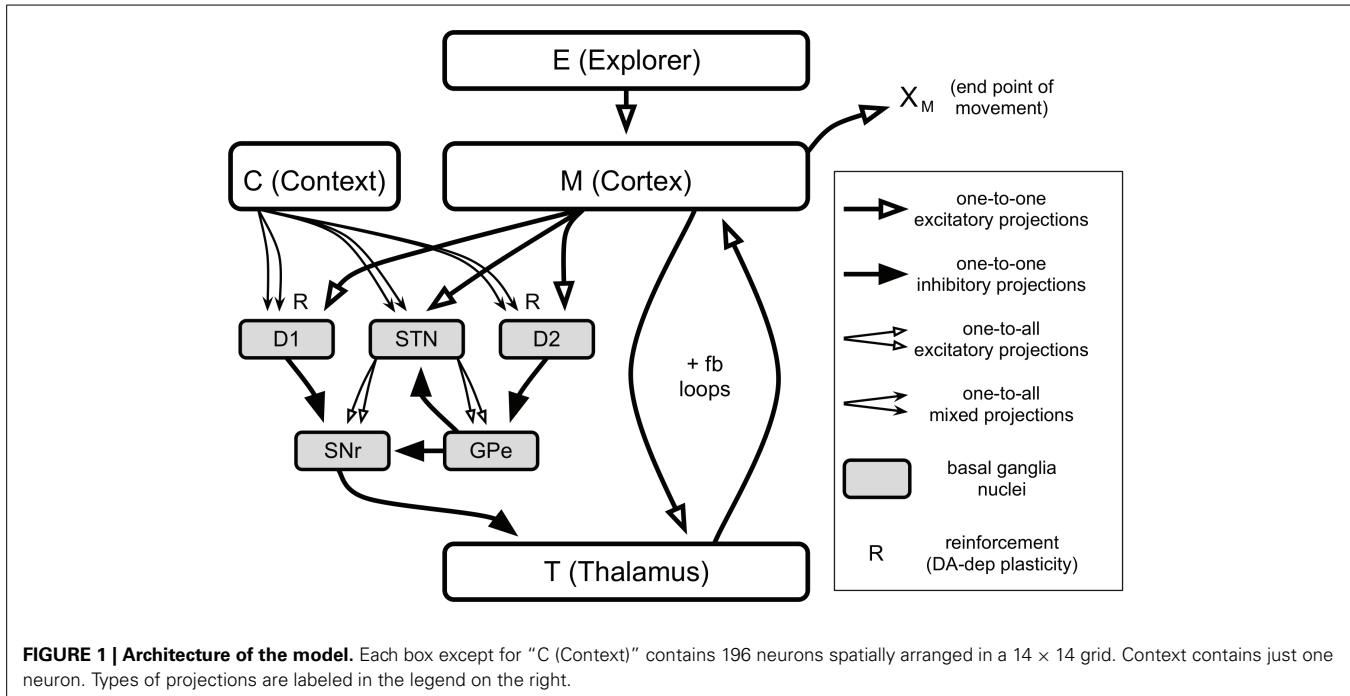
We capture features of this behavior with a hand-crafted function describing, for a decision to move to a particular spatial location, the evolution of activity for every neuron in the Explorer. Early in the process, all neurons are weakly-excited with low activation levels. Neural activity evolves such that, as confidence in a particular movement increases, so does the corresponding neuron activity. The activities of other neurons increase to a lesser degree. An example of this behavior is shown in **Figure 2**; it is described in greater detail in the next paragraph and in the Supplementary section.

For each movement, a particular neuron in Explorer, labeled G_{exp} , is chosen. If we suppose that sophisticated cognitive mechanisms are not devoted to movement selection, G_{exp} is chosen randomly. The activity of the neuron corresponding to G_{exp} increases linearly to one (green line in **Figure 2**). The activities of surrounding neurons change according to a Gaussian-like function centered at G_{exp} . They first increase and then decrease; those furthest from G_{exp} increase by a small amount and then quickly decrease to zero, while those closer to G_{exp} increase by a larger amount and decrease at a later time point to zero. The pattern of activity such that the activity of neuron G_{exp} is one and the activities of all other neurons are at zero is held for brief time, and then the activities of all neurons are set to zero. This evolution takes T_E time steps, which is the number of time steps in a trial.

If, in contrast, we assume sophisticated cognitive mechanisms do influence movement selection, G_{exp} is chosen in order to reflect that strategy, e.g., according to some heuristic search such as a spiral pattern or quadrant-by-quadrant search. In this paper we examine behavior that results when cognitive mechanisms do not influence movement selection as well as behavior that results from a simple pattern, as described in the subsection “Biasing of behavior.”

2.2. CORTEX AND THALAMUS

“Cortex” represents cortical areas that encode high-level movement plans such as reaching or pointing to a location (Anderson and Buneo, 2002). In our model, the spatial location of a neuron in Cortex corresponds to a target spatial location in the workspace, or movement end-point, to which to reach. Cortex (M) receives excitatory projections from Explorer and



Thalamus (T) which preserve channel identity; that is, the neurons representing a given channel in Explorer and Thalamus project to the corresponding neuron in Cortex. In turn, Thalamus receives channel-wise excitatory projections from Cortex, and channel-wise inhibitory projections from SNr (a nucleus of the BG called the substantia nigra pars reticulata). Cortex and Thalamus therefore form a positive feedback loop referred to as a *Cortex-Thalamus loop*, for each channel which is excited by the corresponding channel in Explorer. The gain of a Cortex-Thalamus loop is modulated by inhibitory projections from SNr neuron to Thalamus (Chambers et al., 2011). When the activity level of an SNr channel is low, the corresponding Thalamus neuron is said to be *disinhibited* and its Cortex-Thalamus loop has a high gain. A Cortex-Thalamus loop with a high gain is more easily-excited by the corresponding Explorer neuron.

2.3. BASAL GANGLIA

The functional properties of BG architecture have been described in detail in prior work (Gurney et al., 2001a,b; Humphries and Gurney, 2002; Redgrave et al., 2011). Briefly, the BG is a subcortical group of brain areas with intrinsic architecture that is well-suited to select one behavioral option among competing options. The BG implement an off-center on-surround excitation pattern: The BG channel i that is most strongly-excited by its cortical “action request” inhibits the corresponding target channel (neuron) in Thalamus the least, while other Thalamus channels $j \neq i$ are further inhibited. Thus, Cortex-Thalamus loop i is most easily-excited by input from Explorer to Cortex, and other Cortex-Thalamus loops $j \neq i$ are harder to excite by input from Explorer to Cortex. These properties are similar in some ways to those of a winner-take-all network between the competing

channels, but additional architectural features of the BG ensure better control of the balance between excitation and inhibition (Gurney et al., 2001a,b). D1 and D2 refer to different populations of neurons (named after the dopamine receptors they predominantly-express) in a nucleus of the BG called the striatum. The pathway comprising D1 and STN (subthalamic nucleus) performs the selection with an off-center on-surround network in which D1 supplies focussed (“central”) inhibition and the STN a diffuse (“surround”) excitation. The pathway through D2 regulates the selection by controlling, though GPe (external segment of the globus pallidus), the excitatory activity of STN (Gurney et al., 2001a,b).

2.4. FROM CORTICAL ACTIVITY TO BEHAVIOR

Movement in this model is a function of the activities of the Cortex neurons. Each neuron with an activation greater than a threshold η “votes” to move to the location represented by its grid location with a strength proportional to its activity (i.e., using a population code, Georgopoulos et al. 1982). In most cases, because of the selection properties of the BG, the activation of only one Cortex neuron rises above η . At each time step t , the target location to which to move, $X_M(t)$, is an average of the locations represented by Cortex neurons with activities above η , weighted by their activities. At each t , if any Cortex neuron is above η , a simple “motor plant” causes a movement from the current position ($x_p(t)$) toward $X_M(t)$ (see Supplementary section for equations). Movement evaluation, and hence any learning, is based only on $x_p(T_E)$, the position at time T_E (the last time step of a trial). Thus, end-point of movement, not movement trajectory, is evaluated in this model.

2.5. BIASING OF BEHAVIOR

Targets are circular areas within the workspace. A target is considered hit when $||x_p(T_E) - X_G|| < \theta_G$, where X_G is the location of the center of target G and θ_G (= 1.1) is the radius. Thus, a movement to the location represented by neuron i that corresponds to the center of the target, or to locations represented by the immediate four neighboring neurons, is within the target’s radius. When a target is hit, behavior is biased so that the model is more likely to make movements to the target. This repetition bias (Redgrave and Gurney, 2006) can be implemented in two ways in this model.

The first way is “BG-mediated biasing,” which is based on dopamine-dependent plasticity at the corticostriatal synapses (Calabresi et al., 2007; Wickens, 2009), and is implemented as a Hebbian-like rule governing plasticity to weights onto striatal D1 and D2 neurons. When the end-point of movement is evaluated (at time T_E of a trial), usually only one neuron (i) in each of Cortex, D1, and D2 have an activity above zero. If the target is hit, the weights from Cortex neuron i to D1 neuron i , Cortex neuron i to D2 neuron i , the Context neuron to D1 neuron i , and the Context neuron to D2 neuron i are increased according to equations of the following form (see Supplementary section for full equations):

$$\Delta w_i = \alpha \beta^{N_k-1} y_{pre} y_{post} (W_{max} - w_i), \quad (1)$$

where w_i is the weight, y_{pre} is the activity of the presynaptic neuron, y_{post} is the activity of the postsynaptic neuron, α is a

step-size, W_{max} (= 1) is the maximum strength of a synapse, β (= 0.825) is a *habituation* term (Marsland, 2009), and N_k is the number of times target k has been hit. If the target is not hit, the weights are decreased. Weights from Cortex to striatum have a lower limit of zero, while weights from Context to striatum have a lower limit of -0.1 . Neurons that have greater afferent weights are more-easily excited than are neurons with lower afferent weights.

Neurons in D1 and D2 that correspond to movements that were reinforced are excited by the Context neuron from the first time step of a trial onward, and neurons that correspond to movements that were not reinforced are weakly inhibited by the Context neuron. (We use negative weights to approximate the inhibitory effects of striatal interneurons, Koos and Tepper 1999; Bolam et al. 2006). Thus, weights from the Context neuron to D1 and D2 represent an *a priori* bias in favor of movements that were reinforced, and against movements that were not reinforced. This bias is context-dependent and, while there is only one context for the results reported in this paper, multiple contexts can be represented by multiple context neurons with similar learning rules. Neurons in D1 and D2 are also excited by Cortex neurons, which, early in a trial, are all weakly-excited by Explorer. Because the projections from Cortex to D1 and D2 are plastic, movements that were reinforced are more-easily excited by Cortex than movements that were not reinforced.

Thus, with BG-mediated biasing, channels corresponding to making a movement to locations that are within the target area are easily-excited by weak inputs from the Explorer after the target has been hit several times. Channels corresponding to movements that do not hit the target are made to be more difficult to excite.

The second way by which repetition bias is implemented in this model is referred to as “Cognitive biasing,” whereby G_{exp} is chosen according to some strategy or pattern. Under cognitive biasing in this paper, the set of neurons in Explorer from which G_{exp} is chosen corresponds to a spatial area, centered around the location of the target, that decreases in size each time the target is hit (we describe this pattern in detail in the Supplementary section). This is a simple hand-crafted form of biasing that mimics a decrease in variation and increase in repetition by “zooming in” on the target as the target is repeatedly hit. It is meant to capture the effects of behavioral biasing as mediated by “sophisticated cognitive” or “intelligent” mechanisms. If there is no Cognitive biasing, G_{exp} is randomly chosen as described earlier.

2.6. MODEL EXPERIMENTS

A model run consists of having the model select movements for 300 trials (where a trial consists of executing one movement). Movements were reinforced (Equation 1) when they hit a particular target. We examined behavior that results from reinforcing one target, two targets simultaneously, and one target and then another. The targets are referred to G_1 , G_{2far} (which is far from G_1), and G_{2near} (which is near G_1). Experiments 1 to 4 were conducted to describe patterns of behavior under simple, “non-intelligent,” BG-mediated biasing and different conditions of reinforcement. Experiment 5 was conducted to describe patterns of behavior under BG biasing, Cognitive biasing, and both.

- **Experiment 1: Single target (G_1):** We ran 50 independent runs of 300 movements during which BG biasing (and not Cognitive biasing) was used to reinforce movements that hit G_1 .
- **Experiment 2: Two simultaneous targets (G_1 and $G_{2\text{far}}$):** We ran 50 independent runs of 300 movements during which BG biasing was used to reinforce movements that hit either G_1 or $G_{2\text{far}}$.
- **Experiment 3: Reinforce G_1 , then $G_{2\text{far}}$, then G_1 again:** We ran 50 independent runs of 900 movements during which BG biasing was used to reinforce movements that hit G_1 for the first 300 movements, then to reinforce movements that hit $G_{2\text{far}}$ (but not those that hit G_1) for the next 300 movements, and then reinforce movements that hit G_1 (but not those that hit $G_{2\text{far}}$) for the final 300 movements.
- **Experiment 4: Reinforce G_1 , then either $G_{2\text{far}}$ or $G_{2\text{near}}$:** We ran 50 independent runs of 600 movements during which BG biasing was used to reinforce movements that hit G_1 for the first 300 movements and then to reinforce $G_{2\text{far}}$ (but not those that hit G_1) for the next 300 movements. We ran another 50 independent runs of 600 movements during which BG biasing was used to reinforce movements that hit G_1 for the first 300 movements and then to reinforce $G_{2\text{near}}$ (but not those that hit G_1) for the next 300 movements.
- **Experiment 5: Different bias conditions:** We ran 50 independent runs of 300 movements during which Cognitive biasing (and not BG biasing) was used to reinforce movements that hit G_1 . We ran another 50 independent runs of 300 movements during which both BG biasing and Cognitive biasing were used to reinforce movements that hit G_1 .

3. RESULTS

3.1. EXPERIMENT 1: SINGLE TARGET (G_1)

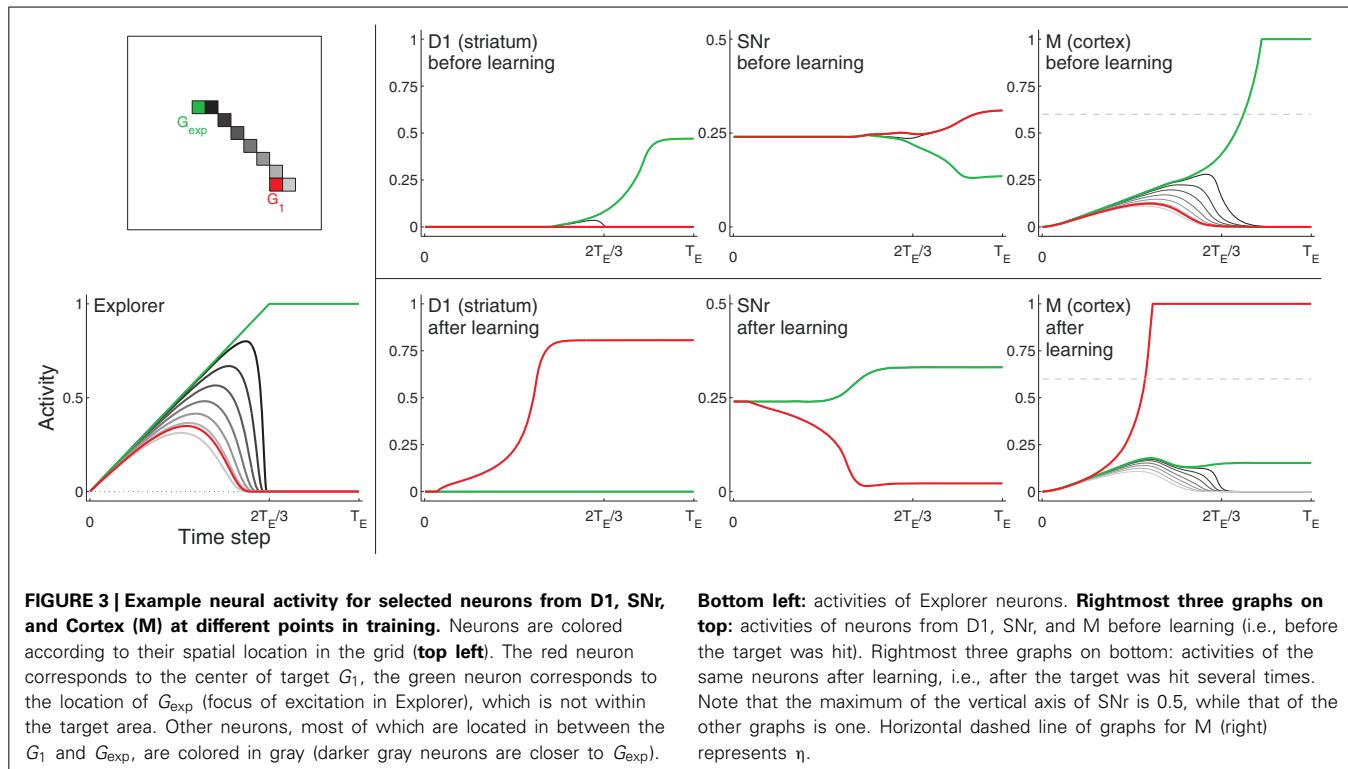
Recall that there are two sources of excitation to the model, as explained in Methods section 2.1: the Context neuron, which projects to D1, D2, and STN; and the Explorer, which projects to Cortex (see also Figure 1). As described in Methods section 2.1, a focus of excitation, G_{exp} , is chosen randomly, and the activities of neurons in the Explorer follow a hand-crafted pattern such that all neurons are weakly-excited initially, but that activity focuses so that only the neuron corresponding to G_{exp} is strongly-excited (see Figure 2). If the weights onto D1 and D2 remain at their initial values, Explorer activity will result in a movement made to the location represented by G_{exp} .

In Experiment 1, there was a single target, G_1 , located in the lower right area of the work space (center of target colored in red in the upper left graph in Figure 3). When the target was first hit, it was because the Explorer happened to choose a G_{exp} that was within θ_G of target center. As described in Methods section 2.5, when the target is hit, the corticostratial weights that project to striatal neurons corresponding to the movement just made are increased (Equation 1). When a target is not hit, the weights decrease. The weight change influences how the BG modulates the gain between Thalamus and Cortex positive feedback loops (Methods sections 2.2 and 2.3), and hence how Cortex responds to excitation from Explorer.

Neural activity

Figure 3 shows selected neuron activity resulting from the same excitation from the Explorer during early movements (“before learning”) and during late movements (“after learning”). Excitation from Explorer is illustrated in the lower left graph, and the color scheme indicating which neuron’s activity is plotted is illustrated in the upper left graph. In this example, activities of neurons corresponding movements made to G_{exp} are plotted in green; those corresponding to the center of the target (G_1) are plotted in red; and those corresponding to a subset of neurons near or between G_{exp} and G_1 are plotted in shades of gray. (Compare with Figure 2 and Methods section 2.1.) G_{exp} is not within the target area. The top row of graphs to the right of the color scheme graph plot neuron activity in striatum D1, neuron activity in SNr, and neuron activity in Cortex in the untrained model. As excitation from Explorer evolved over time, Cortex neurons increased accordingly due to the direct one-to-one projections from Explorer to Cortex and positive feedback loops with Thalamus (as described in Methods section 2.2). Cortex activity directly excited striatal neurons due to direct one-to-one projections to striatum D1 and striatum D2 (as described in Methods section 2.3). In this case, striatal neurons corresponding to G_{exp} increased in activity. Because no learning has occurred yet, Context did not bias activity in striatum as all projections from Context to striatum remained at zero. Intra-BG processing (described in Methods section 2.3) resulted in a decrease in activity of SNr neuron corresponding to G_{exp} , and an increase in all other SNr neurons. This disinhibited the Thalamus neuron corresponding to G_{exp} , increasing the gain on the positive feedback loop with Cortex neuron corresponding to G_{exp} , thus allowing it to increase in activity even more. In addition, the increased activity of all other SNr neurons further decreased the positive feedback gain between other Cortex-Thalamus neuron pairs (Chambers et al., 2011). In this example, weights into D1 and D2 have not undergone any changes, i.e., the target has not been hit, so there is no biasing from Context. Thus, the BG facilitated the selection of the movement suggested by Explorer (move to location G_{exp}) and inhibited the selection of other movements.

After the target had been hit many times, the weights from Context to striatal neurons D1 and D2, and from Cortex to D1 and D2, that correspond to movements made to a location within the target zone (in this example, the center of G_1) increased (as described in Methods section 2.5 and Equation 1), and the weights to all others decreased by a small amount. Neuron activity in response to the same excitation from Explorer after learning is illustrated in the bottom, right most three graphs of Figure 3. Neurons that correspond to G_1 (plotted in red) are referred to as s_G . Because weights from Context to s_G in D1 and D2 have increased, the activity of neuron s_G in D1 and D2 increased faster due to excitation from Cortex than did that of other neurons, including that of neurons that correspond to movements made to G_{exp} . This caused a decrease in the activity of SNr neuron s_G and an increase in the gain of the corresponding Cortex-Thalamus positive feedback loop (described in Methods section 2.2). Hence, the weak excitation to Cortex neuron s_G at the beginning of a movement period was sufficient to initiate a positive feedback process between the corresponding neuron s_G in Cortex



and Thalamus, causing more excitation to neuron s_G in D1 and D2, even further disinhibition of the feedback loop, and further inhibition of the loops of other neurons. BG-mediated bias was in favor of movements toward G_1 , implemented by an increase in weights from Context and Cortex to the neurons in D1 and D2 that correspond to a movement to G_1 (Equation 1). Thus, Cortex neuron s_G increased above η and movement was made to the location corresponding to G_1 , even though the Explorer more-strongly excited neurons corresponding to movements made to G_{exp} .

Movement redistribution under contextual bias

The biasing of activity within the BG, BG's regulation of Cortex-Thalamus loop excitability, and the gradual focusing of excitation from Explorer to Cortex, comprise simple mechanisms that results in a seemingly “intelligent” structured transition from variability to repetition. After the target had been hit by chance a few times, weights from Context to neurons s_G in D1 and D2, and weights from neuron s_G in Cortex to neurons s_G in D1 and D2, were increased a little (Equation 1). When Explorer later chooses G_{exp} near G_1 , the resulting relatively high excitation to Cortex neuron s_G , combined with the increased gain at Cortex-Thalamus loop s_G and decreased gain to other loops, excited Cortex neuron s_G while preventing other Cortex neurons from increasing past η . Thus, a movement to the target was made when Explorer chose G_{exp} near G_1 : the target was hit with an increased likelihood, and movements to areas near the target were made with a decreased likelihood. We refer to this pattern as a “bias zone,” centered at G_1 , that increases in size the more often the target is hit.

Figure 4 shows how the bias zone increases as the number of times the target has been hit increases. In order to produce this figure, the model was run with G_{exp} set to G_1 for a set number of times. Then, learning was turned off and model response for G_{exp} set to each possible location was examined. Each graph in **Figure 4** plots the location of G_{exp} in the workspace: green dots indicate locations of G_{exp} that result in movements made to those locations; red dots indicate locations of G_{exp} that result in movements made to locations within the target area (red circle). The title of each graph indicates how many times G_{exp} was set to G_1 before response was examined. The expansion of the bias zone determines an “intelligent-looking” structured transition from variation to repetition in that it follows a non-random pattern.

For the purposes of this paper, model behavior is considered to be well-learned when a “streak” of hitting the target with ten consecutive movements is achieved. **Figure 5**, top left, plots the proportion of 50 runs that achieved this streak by various points of experience. About 40% reached it by 100 movements, and almost 80% reached it by 300 movements. A little over 20% did not achieve it by 300 movements. **Figure 5**, bottom left, plots the proportion of 50 runs that hit the target as a function of movement number. The proportion reaches about 0.8 by movement number 300.

Figure 5, right, plots, for each movement across the 50 runs, the distance between the movement and G_1 as a function of movement number. The distance of movements that hit G_1 are plotted in red (and are all at zero). As movement number increases, the density of movements near G_1 but that did not hit G_1 decreases at a faster rate than the density of movements far from G_1 . This pattern is due to the expanding bias zone (**Figure 4**). We develop

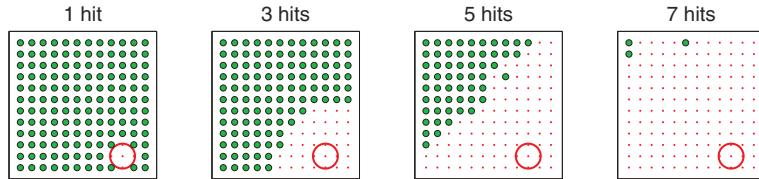


FIGURE 4 | Illustration of the “bias zone” effect. In each graph, the target was first hit N times (labeled at the top of the graph). Then, learning was turned off and movement for each possible value of G_{exp} was evaluated. Each dot represents the spatial location corresponding to G_{exp} . Large green dots

represent locations of G_{exp} that resulted in movements that hit the location corresponding to G_{exp} . Small red dots represent locations of G_{exp} that, because of biasing implemented by weights onto D1 and D2, resulted in movements that hit the target (represented by the red circle in the lower right).

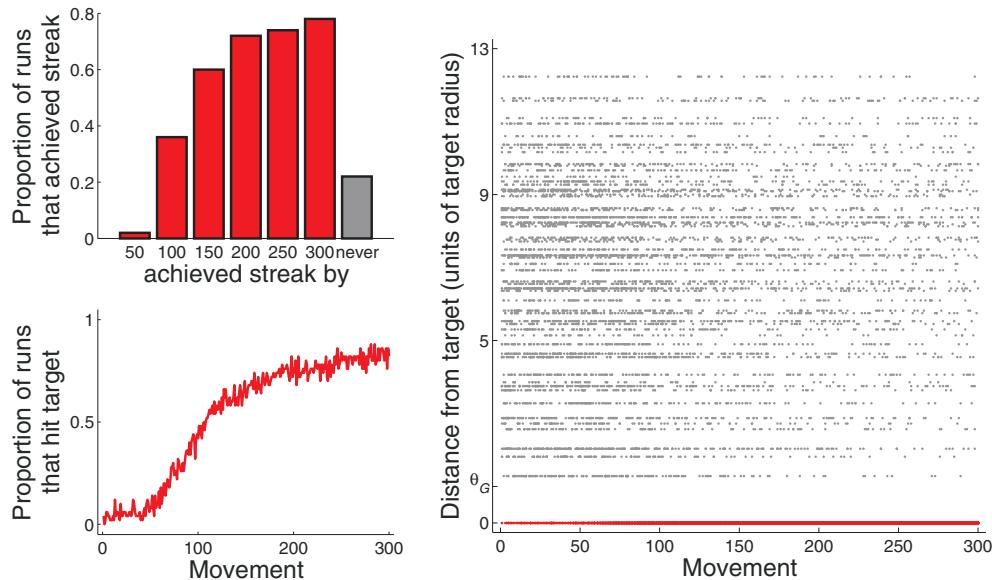


FIGURE 5 | Performance across all 50 runs for Experiment 1: A single target (and only BG biasing). **Top left:** proportion of runs that achieved streak of hitting target ten consecutive times by the movement 50, 100, 150, ..., or 300. Note that the bar graphs are cumulative. **Bottom**

left: proportion of runs that hit target as a function of movement number. **Right:** Distance from the center of the target (in units of target radius) of each movement from all 50 runs. That for movements that hit the target are drawn in red and are at value 0 of the vertical axis.

a method for quantifying this pattern in the section describing results of Experiment 5 (and in the Supplementary section). Experiment 4 describes behavior in a more complicated task that results from this pattern.

Effect of cortical noise on model performance

The capability of the model to bias movements toward G_1 is due in part to the pattern of excitation from Explorer to Cortex (**Figure 2**), which weakly-excites all Cortex neurons by very similar amounts early in a trial. This suggests that model performance may be sensitive to unpredicted deviations from this pattern. To investigate this, we ran simulations in which signal-dependent noise (Harris and Wolpert, 1998) was added to Cortex neurons (which project to the BG and Thalamus, and from which movement is determined). In particular, at each time step: $y \leftarrow [y + y N(0, \sigma)]_0^1$, where y is the output activity of a Cortex neuron, $N(0, \sigma)$ refers to a number drawn randomly from a zero-mean Gaussian distribution with standard deviation σ , and $[x]_0^1$ returns 0 if $x < 0$, 1 if $x > 1$, and x otherwise. The proportion

of the last 30 movements of all runs under a particular noise condition that were made to G_1 were 0.82, 0.64, 0.53, and 0.20 for σ levels of 0 (no noise), 0.1, 0.3, and 0.5, respectively. Thus, the model was able to learn to repeatedly hit G_1 if a low to moderate level of noise was added to Cortex neuron activity, but performance dropped off with high levels of noise. **Figure 6** illustrates, in a manner similar to **Figure 3**, example model neuron activity for a model run with $\sigma = 0.1$. The rest of the simulations in this paper were run with no noise.

3.2. EXPERIMENT 2: TWO SIMULTANEOUS TARGETS (G_1 AND $G_{2\text{far}}$)

Movements that hit either of two targets, G_1 (lower right of the workspace) or $G_{2\text{far}}$ (upper left) (red and blue circles, respectively, in **Figure 7**), were reinforced according to Equation 1. However, the habituation term differentiated them. (The habituation term is β^{N_k-1} in Equation 1, where N_k is the number of times target k has been hit and $\beta = 0.825$.) For example, even if G_1 was hit many times, at the first time $G_{2\text{far}}$ was hit, it was a novel event and thus the corresponding weights increased by a large amount.

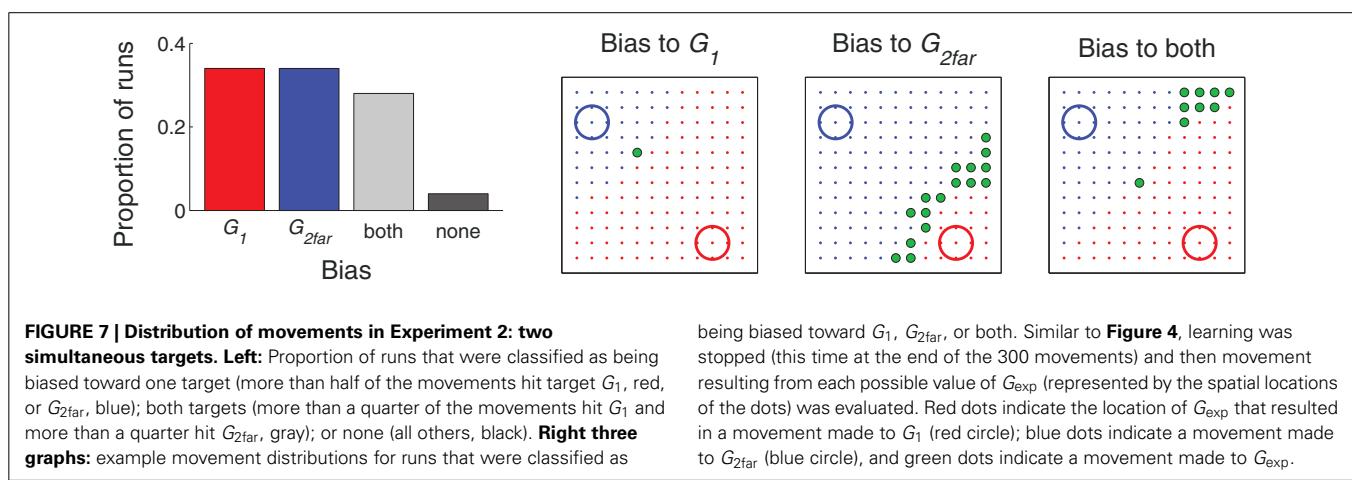
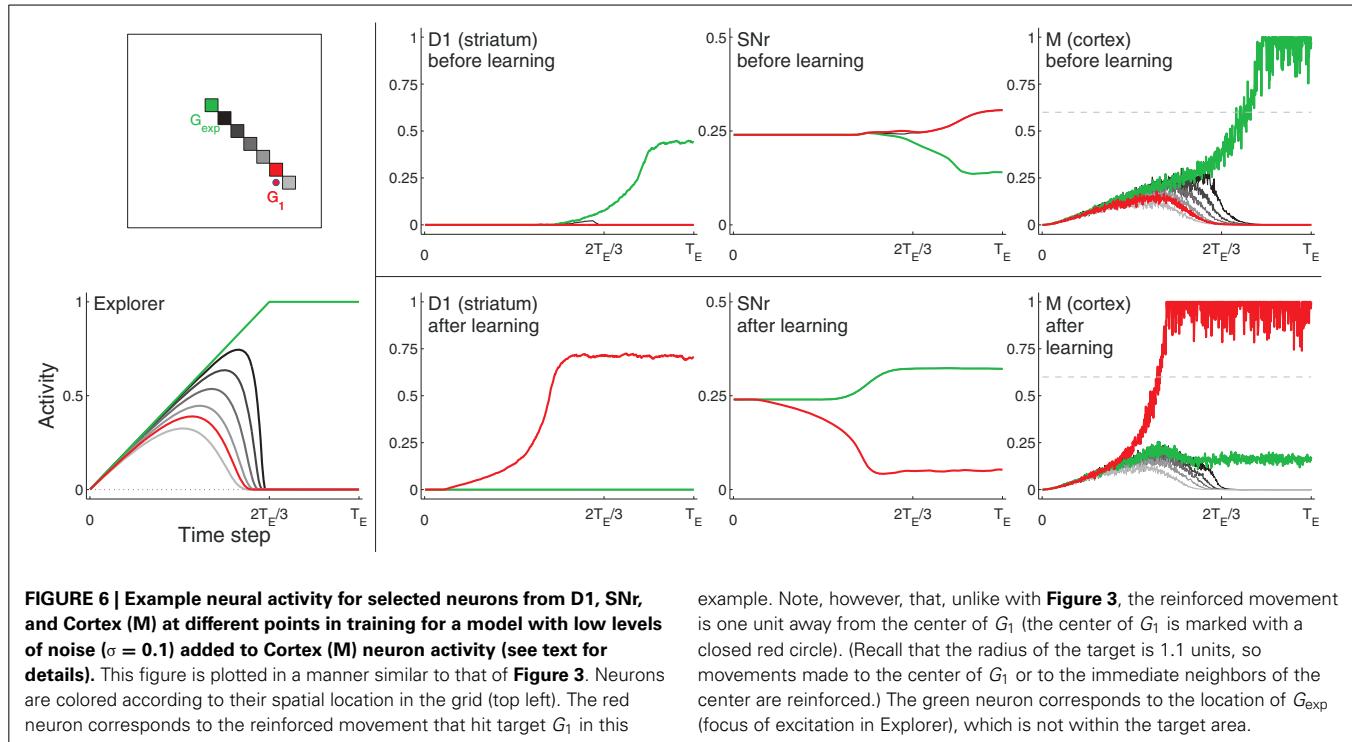


Figure 7, left, plots the proportion of runs that were classified as either biased toward one of the targets, distributed between the two targets, or did not find a target (see figure caption for details on the classification criteria). While behavior in a majority of the runs was biased to a single target (e.g., middle two graphs of **Figure 7**), the model was capable of distributing movements to both targets (e.g., **Figure 7**, right). For runs which were biased to just one target, only a G_{exp} very near the un-preferred target produced a movement to that target.

3.3. EXPERIMENT 3: REINFORCE G_1 , THEN $G_{2\text{far}}$, THEN G_1 AGAIN

The use of experience-based learning rules—weight modification (Equation 1) is dependent on actual behavior—and a habituation term leads to a type of memory that can influence subsequent

behavior in a changing environment. This is illustrated with experiments in which only movements to G_1 are reinforced for 300 movements, then only movements to $G_{2\text{far}}$ are reinforced (at which point the habituation term for G_1 is reset), and then only movements to G_1 are reinforced again. As shown in **Figure 8**, top row, which plots the proportion of runs that hit each target as a function of movement number, the reacquisition of G_1 (movements 601–900) occurred faster than the initial acquisition (movements 1–300) of G_1 .

The enhanced acquisition is because corticostriatal weights corresponding to movements toward G_1 , illustrated in red in **Figure 8**, bottom row, increased to a stable value (of about 0.2 in the figure) during first acquisition. (The habituation prevents it from increasing any more after the target had been

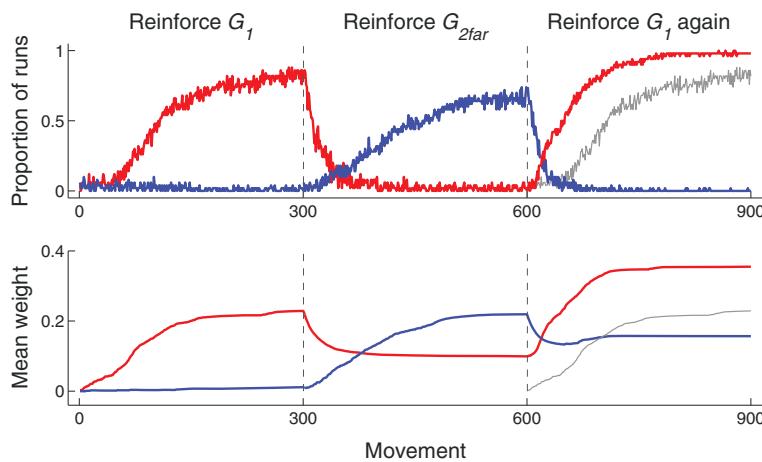


FIGURE 8 | Time-course of behavior and corticostriatal weights for Experiment 3: reinforce G_1 (movements 1–300), then $G_{2\text{far}}$ (301–600), then G_1 again (601–900). **Top:** proportion of runs that hit G_1 (red) or $G_{2\text{far}}$ (blue) as a function of movement number. The proportion of runs that hit G_1 during movements 1–300 are redrawn at horizontal positions 601–900 as a gray line for comparison of performance between initial acquisition (movements 1–300) and reacquisition (movements 601–900) of G_1 . **Bottom:** Mean (across runs) weight from Context neuron to the D1 neuron corresponding to most movements that hit G_1 (red) or $G_{2\text{far}}$ (blue) for that

particular run. The D1 neuron that corresponded to most movements that hit each target was determined by finding the maximum weight from Context to D1 neurons at the end of each 300 movement segment. Because several movements can hit each target, only runs in which the same D1 neuron was selected at movement 300 and movement 900 (i.e., for movements that hit G_1) were included (16 out of 50 runs were excluded). That for weights from Context neuron to D2 neurons followed a similar pattern and are not plotted. Similar to the graphs in the top row, mean weight during movements 1–300 are plotted again at movements 601–900 in gray for comparison purposes.

repeatedly hit.) During movements 301–600, $G_{2\text{far}}$ was reinforced (and G_1 was no longer reinforced). The model continued to move to G_1 early in the second set of movements, but, because G_1 was no longer reinforced, the corresponding weights decreased. As the weights decreased, the bias zone around G_1 decreased and the model was free to move to other locations, including toward $G_{2\text{far}}$. As a new bias zone, now centered on $G_{2\text{far}}$, was established, the model stopped moving to G_1 . Because movements toward G_1 were no longer made, weights associated with moving to G_1 ceased to decrease. When movements to G_1 were reinforced again, those weights were already above zero and thus G_1 was reacquired faster than it was initially acquired. In addition, due to resetting the habituation term, the weights increased to a greater value than the previous high value.

This pattern of activity provides a simple mechanism that can be used to partially explain the findings that practice sessions that are separated in time lead to enhanced acquisition and performance compared to practice sessions that are massed together (Ammons, 1950; Baddeley and Longman, 1978) (though such effects do not necessarily apply to all types of tasks, e.g., Lee and Genovese 1989).

3.4. EXPERIMENT 4: REINFORCE G_1 , THEN EITHER $G_{2\text{far}}$ OR $G_{2\text{near}}$

When one target is reinforced for a period of time, and then another is reinforced instead, how well the second reinforced target is acquired depends on its proximity to the first target. This is illustrated by comparing the results of experiments in which the second target ($G_{2\text{far}}$, blue in Figure 9) was far from the first one with those in which the second target ($G_{2\text{near}}$, purple) was near the first one. Figure 9 plots

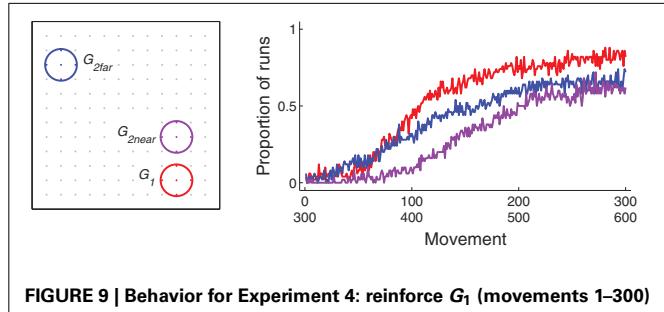


FIGURE 9 | Behavior for Experiment 4: reinforce G_1 (movements 1–300) and then either $G_{2\text{far}}$ or $G_{2\text{near}}$ (301–600). **Left:** locations of the three targets in the workspace (dots indicate locations corresponding to possible values of G_{exp} , colored gray if those locations do not lie within a target area. **Right:** proportion of runs that hit G_1 for movements 1–300 or $G_{2\text{far}}$ or $G_{2\text{near}}$ for movements 301–600.

the proportion of runs for which the first and second targets were hit as a function of movement for the different sets. The first target (G_1 , red) was acquired the fastest. The far second target ($G_{2\text{far}}$) was acquired faster than the near second target ($G_{2\text{near}}$).

The discrepancy between acquiring the second targets is explained by the bias zone. A well-learned model has corticostriatal weights such that the bias zone is large. When the bias zone is centered around G_1 , un-reinforced movements to G_1 must happen in order for weights to decrease, after which the bias zone shrinks and movements to other locations can be made. Movements to locations far from G_1 are available earlier than movements to locations near G_1 as the bias zone shrinks. Thus, a second target far from G_1 will be more-easily acquired than a second target near G_1 .

3.5. EXPERIMENT 5: MOVEMENT REDISTRIBUTION UNDER DIFFERENT BIAS CONDITIONS

As movements made to a target increase, movements made to other locations must decrease: movements are redistributed over the workspace. The previous sections focused on movement redistribution in our model with only BG-mediated biasing (Equation 1). Here we describe metrics of movement redistribution that will allow us to compare how movements are redistributed under different bias conditions. We focus on model runs in which only movements made to one target (G_1) were reinforced.

Redistribution metric

The expanding bias zone (Results section 3.1 and **Figure 4**) that results from BG-mediated biasing results in a pattern of behavior such that movements made near, but not at, the target decrease in likelihood earlier than movements made far from the target. For each run, we quantify the rate of decrease as a function of distance from target. Briefly (see **Figure 10**), movements that did not hit the target were coarsely categorized into three temporal chunks and three spatial zones (vertical and horizontal lines, respectively, in **Figure 10**). Temporal chunk one includes the first 100 movements; temporal chunk two includes the second 100 movements; and temporal chunk three includes the last 100 movements. Recalling that θ_G is target radius and letting dX be the distance of a movement from target center, the spatial zones are 1) $\theta_G < dX \leq 5\theta_G$ (green points in **Figure 10**), 2) $5\theta_G < dX \leq 9\theta_G$ (blue), and 3) $9\theta_G < dX$ (black). The number of movements that fell into spatial zone i from temporal chunks 1 to 2 to 3 was fit to an equation of the form $e^{b_i(j-1)}$, where j refers to temporal chunk. The rate of decrease of the number movements was quantified by the parameter b_i . A more negative b_i indicates a greater rate of decrease (see the Supplementary section for more details).

Movement redistribution across different bias conditions

Figure 10, top row, graphs movement distance from target as a function of movement number for three sample runs under BG-mediated bias (these graphs are similar to **Figure 5**, right). In all three cases, $b_1 < b_2 < b_3$, i.e., the rate of decrease of movements made near but not at the target is greater than that of movements made far from the target. This is in line with the behavioral pattern we would expect given the expanding bias zone (**Figure 4**) that results from BG-mediated biasing. Regarding the specific sample runs in **Figure 10**, top row, the rate of decrease of movements from the first sample run that fell within zone one is greater than that of the second sample run, which is greater than that of the third sample run. This, also, is reflected in the b metrics.

The same process was used to determine b metrics for models that biased movement selection with different mechanisms. Recall from Methods section 2.5 that, if there is no “Cognitive bias,” movements suggested by the Explorer (G_{exp}) were randomly selected from a uniform distribution over all possible movements. Under the Cognitive bias scheme (described in Methods section 2.5 and the Supplementary section), every time the target is hit, the set of possible movements from which G_{exp} is selected decreases: movements further from target center are

removed from the set earlier than movements closer to target center. Movement redistribution under a Cognitive bias thus follows a trend opposite that under BG-mediated bias: $b_1 > b_2 > b_3$ (**Figure 10**, bottom row).

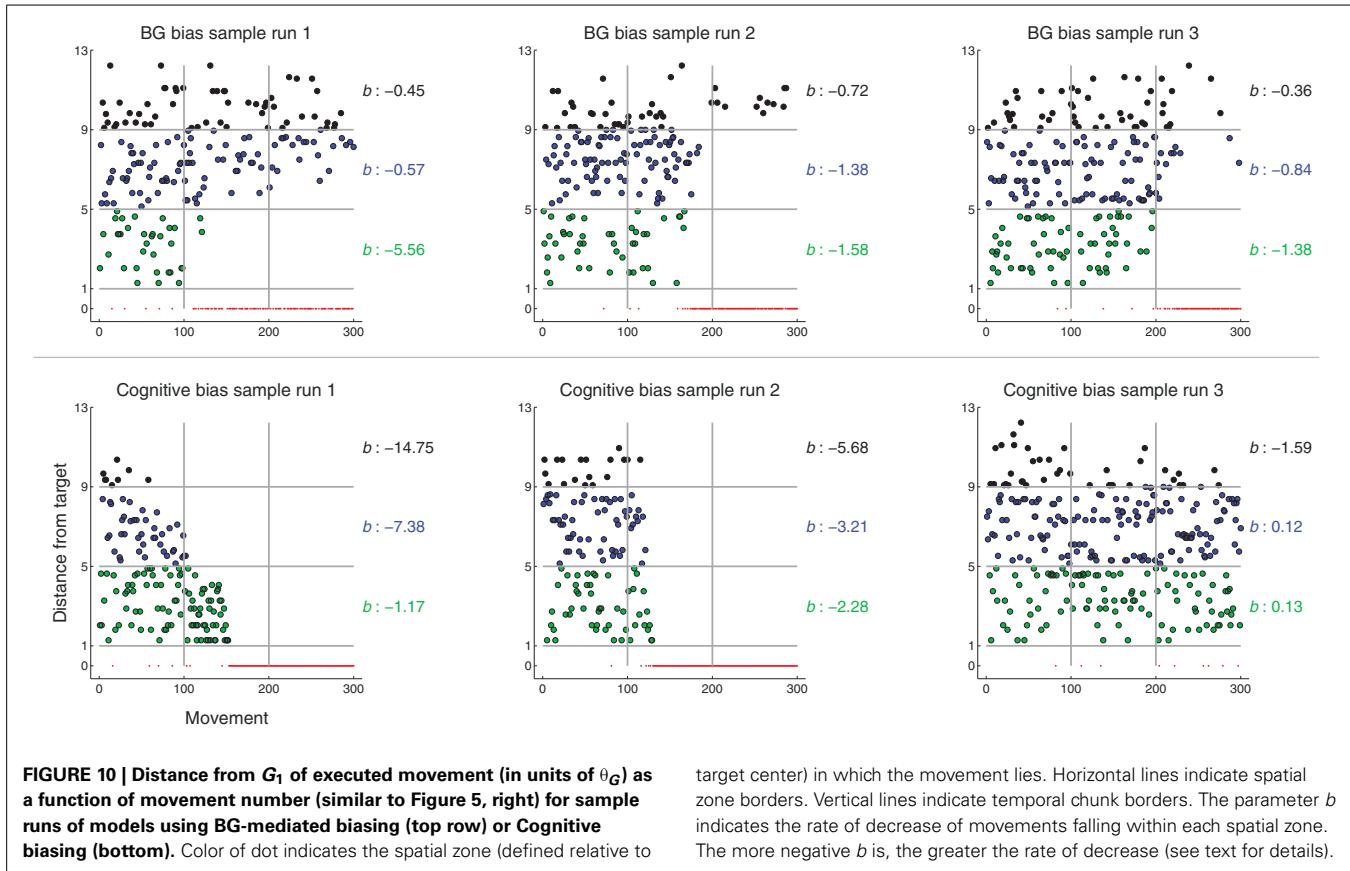
For a given run of a model using BG-mediated bias, b for spatial zones closer to the target should be more negative than b for zones farther from the target. Thus, we expect $b_3 - b_2 > 0$ and $b_2 - b_1 > 0$ in models using BG-mediated bias. Models using the Cognitive bias should exhibit opposite behavior: $b_3 - b_2 < 0$ and $b_2 - b_1 < 0$. The differences should be zero if the transition from variation to repetition does not follow a structured pattern (i.e., the frequency of movements to non-target areas decreases uniformly).

Figure 11 plots the distribution of pair-wise (by run) differences $b_3 - b_2$ (right column, black) and $b_2 - b_1$ (left column, blue) of model runs using different bias conditions (arranged by row). The means of the distributions were also tested against the null hypothesis that they are zero (single sample one-tailed t -tests). The distributions of the pair-wise differences for models using a BG bias (top row) were positive; that for models using a Cognitive bias (bottom) were negative; and that for using both biasing mechanisms (middle) were also negative (though visual inspection suggests that the Cognitive bias condition has more extreme negative pair-wise differences than does the combined bias condition). Thus, this analysis was able to capture the general trends that were seen in the different bias conditions of the model.

4. DISCUSSION

As described in a recent theory of *action discovery* (Redgrave and Gurney, 2006; Redgrave et al., 2008, 2011, 2013; Gurney et al., 2013), when an unexpected sensory event occurs, animals transition from executing a variety of movements to repeating movements that may have caused the event. A transition from variation to repetition often follows non-random, structured patterns that may be explained with sophisticated cognitive mechanisms (e.g., Dearden et al. 1998; Dimitrakakis 2006; Simsek and Barto 2006). However, in action discovery, simple non-cognitive mechanisms involving dopamine modulation of basal ganglia (BG) activity are thought to play a prominent role in behavioral biasing. In this paper we use a biologically-plausible computational model to demonstrate that a structured transition from variation to repetition can emerge from processing within such simple mechanisms. Such behavior is due to the following features on which our model, unlike most previous models of BG function, focuses: (i) the BG does not bias behavior directly, but modulates cortical response to excitation (Chevalier and Deniau, 1990; Mink, 1996; Humphries and Gurney, 2002; Cohen and Frank, 2009; Redgrave et al., 2011; Baldassarre et al., 2013); (ii) excitation to cortex follows a pattern that evolves from weakly exciting all neurons to strongly exciting only one neuron (Britten et al., 1992; Platt and Glimcher, 1999; Huk and Shadlen, 2005; Bogacz et al., 2006; Gold and Shadlen, 2007; Lepora et al., 2012). By including these features in our model, we show that sophisticated cognitive mechanisms may not always be necessary to develop a structured transition from variation to repetition.

In our model, movements occur by selecting an end-point (spatial location) to which to move. Movements that terminated

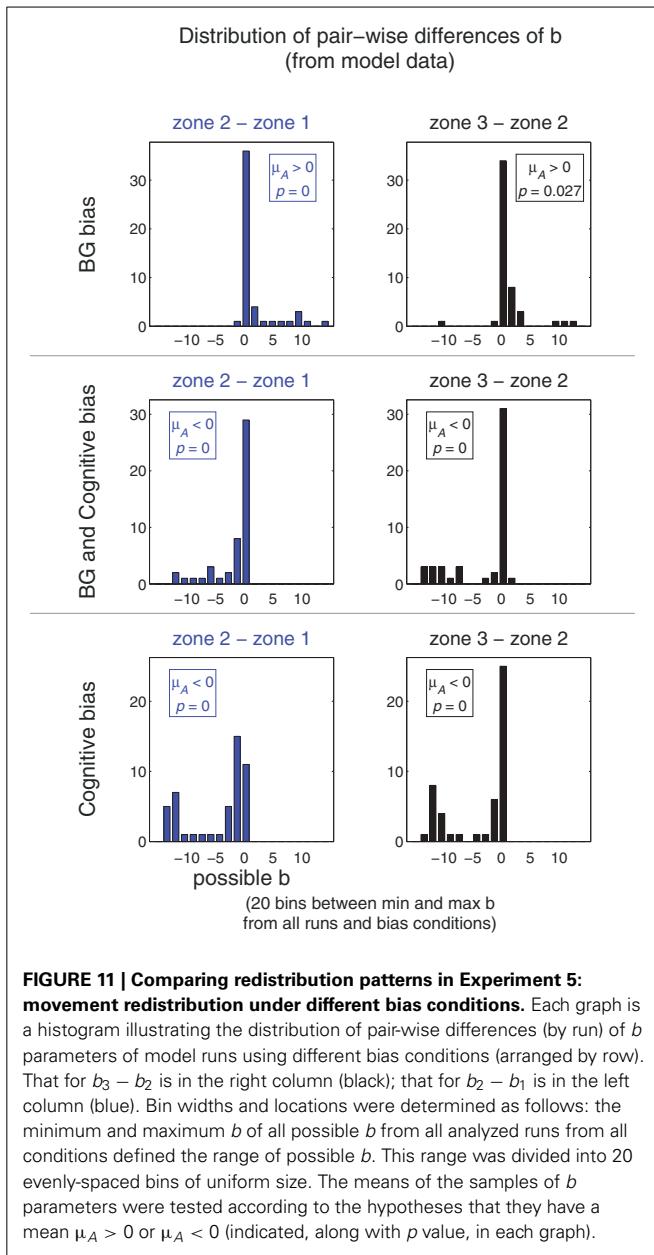


in a target area were reinforced so that the selection of such end-points increased in frequency. The transition from executing a variety of movements to executing just the reinforced movements followed a structured pattern: as end-points at the target location increased in frequency, end-points near, but not at, the target location decreased in frequency at a greater rate than end-points far from the target. We refer to the area around the target area in which end-point frequency decreased as a “bias zone” (**Figures 4, 10, top**), and the bias zone increased in size as the target was repeatedly hit. The graded shift from variation (a small bias zone) to repetition (a large bias zone) allows for the discovery of a second target area in some cases (**Figure 7**), and also results in specific patterns of behavior if the target area moves (**Figures 8, 9**).

In addition, in action discovery, phasic DA activity in response to achievement of the outcome (e.g., hitting the reinforced target area) decreases as associative brain areas learn to predict the outcome’s occurrence (Redgrave and Gurney, 2006; Redgrave et al., 2008, 2011, 2013; Gurney et al., 2013; Mirolli et al., 2013). This may be thought of as a type of intrinsic motivation (IM) in that the outcome need not have hedonic value in order to be reinforcing (Oudeyer and Kaplan, 2007; Baldassarre, 2011; Barto, 2013; Barto et al., 2013; Gottlieb et al., 2013; Gurney et al., 2013). The type of IM in action discovery is best described as some combination of novelty and surprise (Barto et al., 2013). A detailed account of exactly how the prediction process may be implemented in the brain is beyond the scope of this paper.

We mimic its effects in our model with a simple habituation mechanism similar that used in neural network models of novelty detection (Marsland, 2009). Here, the reinforcing effects of an outcome with which the model has little recent experience is greater than the reinforcing effects of an outcome with which the model has much recent experience. The habituation term (β^{N_k-1} in equation 1) influences behavioral patterns, particularly in tasks in which more than one target area is reinforced (**Figure 7**) or the target area changes (**Figures 8, 9**). Unlike the reward prediction error hypothesis of phasic DA neuron activity (Houk et al., 1995; Schultz et al., 1997), habituation is a mechanism that does not rely on extrinsic motivation by which phasic DA neuron activity, and hence rate of change of the rate of cortico-striatal plasticity, decreases with continued occurrences of the outcome.

We also implement models in which a structured transition from variation to repetition is that which would be expected if one type of more sophisticated mechanism (“Cognitive biasing”) is in effect. The pattern of behavior (**Figure 10, bottom**) is then different than that of BG-only biasing. Finally, we have devised a method for capturing such differences with quantitative measures (**Figures 10, 11**) which will allow us to make contact with future behavioral experiments investigating how different brain areas contribute to biasing behavior in tasks similar to model tasks. In continuing work, we are devising such behavioral experiments. Preliminary results suggest that our quantitative measure will allow us to compare the effects of different biasing mechanisms



by examining behavior from different systems (e.g., model versus human), different workspaces, different target sizes, and different target locations, etc. Possible mechanisms by which to isolate different brain mechanisms include explicit instructions, use of different stimuli (Thirkettle et al., 2013b), or use of distractor tasks (Stocco et al., 2009).

As with any computational model of brain systems, the mechanisms described in this paper should be viewed as being a part of a complex system of interacting parts. We've isolated the effects of the specific mechanisms we've investigated in order to demonstrate how a structured transition from variation to repetition can emerge from those mechanisms. In the next subsection we discuss the implications of some of these choices in greater detail and how to expand on them to include more sophisticated systems.

4.1. A MULTI-STAGE SELECTION PROCESS

Recall that, for each movement in our model, the pattern of excitation from “Explorer” to “Cortex” evolves from weakly-exciting all neurons to strongly-exciting one neuron (referred to as G_{exp} , the focus of excitation). The weak excitation of all neurons early in the evolution allows for corticostriatal plasticity to bias behavior. Behavior can also be biased by the choice of G_{exp} , the effects of which are greater later in the evolution. Thus, the evolving excitation pattern from Explorer to Cortex allows for a multi-stage selection process. We expand on these points below.

Through corticostriatal plasticity and BG selection mechanisms, Cortex neurons that are only weakly excited during the early stages of excitation from Explorer can increase in activity at a greater rate than other Cortex neurons. BG selection mechanisms also enable these neurons to suppress the responses of other Cortex neurons to subsequent strong excitation (e.g., Figure 3). The expanding bias zone (described in Results section 3.1 and Figure 4) that is seen in models using BG-mediated biasing emerges from the pattern of excitation from Explorer to Cortex. Because the model task was a spatial reaching task, a topographic representation was used that revealed an apparent dependency between movements: neurons in Explorer near the focus of excitation (G_{exp}) were excited more than neurons far from the focus.

However, a different pattern may be revealed in other types of tasks. In general, the pattern of activity is likely to be influenced by perceptual processing of sensory information. For example, the theory of affordances (Gibson, 1977, 1986) suggests that the perception of objects preferentially primes neurons that correspond to actions that can operate on those objects, e.g., the perception of a mug would prime a grasping action. Thus, the pattern of excitation in these conditions would preferentially excite those neurons, and excitation may follow a pattern that is different than the one used in this paper. Because BG modulates how Cortex responds to excitation rather than directly-exciting movements, any behavioral pattern controlled by BG-mediated biasing would depend on the pattern of excitation to Cortex. Thus, different patterns of exploration, and different patterns of a structured transition from variation to repetition, would be observed in different environments and tasks.

We envision that more sophisticated mechanisms (e.g., our Cognitive biasing) can be expressed in our model in the later part of the evolving excitation pattern of the Explorer, i.e., in how G_{exp} is chosen. One such mechanism may search the workspace in a way that is more intelligent than random, such as a spiral or raster-like search pattern that does not repeat itself until all possible movements have been executed. The choice of G_{exp} could also be adaptive, including using mechanisms by which a transition from variation to repetition is governed by mechanisms based on measures of optimality, uncertainty, or other task-related variables (Dearden et al., 1998; Daw et al., 2006; Dimitrakakis, 2006; Simsek and Barto, 2006; Cohen et al., 2007).

Thus, the early part of the evolving excitation pattern from Explorer to Cortex comprises weak excitation that is influenced by perception of the environment (e.g., affordances or, in our model, possible movement locations) or simple mechanisms. The later

part of the evolution allows for more complicated mechanisms that may require more processing time to also influence behavior. We have focused mostly on simple mechanisms in this paper, but the evolving pattern of excitation can be used to implement proposed theories that focus on multiple influences on behavior, e.g., Kawato (1990); Rosenstein and Barto (2004); Daw et al. (2005); Shah and Barto (2009).

4.2. ACTION DISCOVERY WITH COMPLICATED BEHAVIORS

There are many types of movements or behaviors that can affect the environment, e.g., making a gesture (regardless of spatial location), manipulating objects in the environment, or making a sequence of movements. In this paper we focused on a simple type of action in which the system, able to select a spatial endpoint of movement, must discover the end-point(s) that delivers an outcome. On a more abstract level, this is similar to “*n*-armed bandit” problems, in which the system must discover which out of a set of *n* actions is followed by the most rewarding consequences in a one-step decision task (e.g., Sutton and Barto 1998). The general process of action discovery (Redgrave and Gurney, 2006; Redgrave et al., 2008, 2011, 2013; Gurney et al., 2013) is also concerned with discovering the temporal and structural components of a complex behavior that affects the environment. These problems are similar to those of temporal and structural credit assignment problems (Minsky, 1961; Sutton, 1984, 1988; Barto, 1985; Sutton and Barto, 1998), which we briefly describe below.

One form of the temporal credit assignment problem is exposed in systems in which a series of actions is required in order to achieve an outcome, and there is great redundancy: a large number of different (but possibly overlapping) sequences can achieve the outcome. How does the agent discover the most direct sequence, i.e., the sequence that uses the fewest actions? This redundancy is often resolved by assigning a cost for each executed action and using optimal control methods to achieve the goal while also minimizing cost (e.g., Sutton and Barto 1998). However, optimal control methods, which are designed to find behavior that minimizes cost according to an arbitrary cost function, may use mechanisms that are more sophisticated and complicated than those thought to underly action discovery. Recent modeling work (Shah and Gurney, 2011; Chersi et al., 2013) has shown that a simpler learning rule that does not incorporate cost per action can discover the most direct sequence of actions in a redundant system. Such behavior remains stable for a period of time, but, if learning is not attenuated, extraneous actions are incorporated with extended experience (Shah and Gurney, 2011).

The structural credit assignment problem is exposed when a system can execute many actions simultaneously and the outcome depends only on the simultaneous execution of a small subset of those. When behavior is composed of several components, and the outcome is contingent on only some of those components, variation allows the animal to determine which components are relevant and to “weed out” the irrelevant components. We have not addressed this problem directly, but previous work on the structural credit assignment problem in RL offers promising directions (Barto and Sutton, 1981; Barto et al., 1981; Barto, 1985; Barto and Anandan, 1985; Gullapalli, 1990).

4.3. CONCLUSION

How biasing causes a transition from variation to repetition so as to converge on the specific movements that cause an outcome is a fundamental problem in the process of action discovery. With a simple model of a restricted aspect of action discovery, which includes neural processing features not included in most other models of BG function, we are able to describe the effects of different types of behavioral biasing. The results reported in this paper describe a first step in understanding the more processes at work in general action discovery.

ACKNOWLEDGMENTS

We are grateful for the efforts of Martin Thirkettle (Open University and University of Sheffield) and Jennifer Tidman (Open University) who, along with Ashvin Shah, developed and ran the preliminary behavioral experiments referred to in the Discussion. In addition, the authors had helpful discussions with Peter Redgrave, Tom Stafford, Jen Lewis, and Nicolas Vautrelle (all at the University of Sheffield), and Patricia Shaw (Aberystwyth University). Finally, we thank anonymous reviewers for their careful reading and comments.

FUNDING

We are grateful for financial support from the European Union’s Seventh Framework Programme grant FP7-ICT-IP-231722 (IM-CLeVeR: Intrinsically Motivated Cumulative Learning Versatile Robots).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <http://www.frontiersin.org/journal/10.3389/fpsyg.2014.00091/abstract>

REFERENCES

- Ammons, R. (1950). Acquisition of motor skill: III. effect of initially distributed practice on rotary pursuit performance. *J. Exp. Psychol.* 40, 777–787. doi: 10.1037/h0061049
- Anderson, R., and Buneo, C. (2002). Intentional maps in posterior parietal cortex. *Ann. Rev. Neurosci.* 25, 189–220. doi: 10.1146/annurev.neuro.25.112701.142922
- Baddeley, A., and Longman, D. (1978). The influence of length and frequency of training session on the rate of learning to type. *Ergonomics* 21, 627–635. doi: 10.1080/00140137808931764
- Baldassarre, G. (2011). “What are intrinsic motivations? A biological perspective,” in *Proceedings of the International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob-2011)*, eds A. Cangelosi, J. Triesch, I. Fasel, K. Rohlfing, F. Nori, P. Y. Oudeyer, et al. (Piscataway, NJ: IEEE), E1–E8.
- Baldassarre, G., Mannella, F., Fiore, V., Redgrave, P., Gurney, K., and Mirolli, M. (2013). Intrinsically motivated action-outcome learning and goal-based action recall: a system-level bio-constrained computational model. *Neural Netw.* 41, 168–187. doi: 10.1016/j.neunet.2012.09.015
- Barto, A. (1985). Learning by statistical cooperation of self-interested neuron-like computing elements. *Hum. Neurobiol.* 4, 229–256.
- Barto, A. (2013). “Intrinsic motivation and reinforcement learning,” in *Intrinsically Motivated Learning in Natural and Artificial Systems*, Chapter 1, eds G. Baldassarre and M. Mirolli (Berlin; Heidelberg: Springer-Verlag), 17–47. doi: 10.1007/978-3-642-32375-1-2
- Barto, A., and Anandan, P. (1985). Pattern-recognizing stochastic learning automata. *IEEE Trans. Syst. Man Cybern.* 15, 360–375. doi: 10.1109/TSMC.1985.6313371
- Barto, A., Mirolli, M., and Baldassarre, G. (2013). Novelty or surprise? *Front. Psychol.* 4:1. doi: 10.3389/fpsyg.2013.00097

- Barto, A., and Sutton, R. (1981). Landmark learning: an illustration of associative search. *Biol. Cybern.* 42, 1–8. doi: 10.1007/BF00335152
- Barto, A., Sutton, R., and Brouwer, P. (1981). Associative search network: a reinforcement learning associative memory. *Biol. Cybern.* 40, 201–211. doi: 10.1007/BF00453370
- Bertsekas, D., and Tsitsiklis, J. (1996). *Neuro-Dynamic Programming*. Belmont, MA: Athena Scientific.
- Bogacz, R., Brown, E., Moehlis, J., Holmes, P., and Cohen, J. (2006). The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. *Psychol. Rev.* 113, 700–765. doi: 10.1037/0033-295X.113.4.700
- Bolam, J., Bergman, H., Graybiel, A., Kimura, M., Plenz, D., Seung, H., et al. (2006). “Group report: microcircuits, molecules, and motivated behavior—microcircuits in the striatum,” in *Microcircuits: The Interface Between Neurons and Global Brain Function*, Dahlem Workshops Reports, Chapter 9, eds S. Grillner and A. Graybiel (Cambridge, MA: MIT Press), 165–190.
- Britten, K., Shadlen, M., Newsome, W., and Movshon, J. (1992). The analysis of visual motion: a comparison of neuronal and psychophysical performance. *J. Neurophysiol.* 12, 4745–4765.
- Calabresi, P., Picconi, B., Tozzi, A., and DiFilippo, M. (2007). Dopamine-mediated regulation of corticostratial synaptic plasticity. *Trends Neurosci.* 30, 211–219. doi: 10.1016/j.tins.2007.03.001
- Chambers, J., Gurney, K., Humphries, M., and Prescott, T. (2011). “Mechanisms of choice in the primate brain: a quick look at positive feedback,” in *Modelling Natural Action Selection*, Chapter 17, eds A. Seth, T. Prescott, and J. Bryson (Cambridge: Cambridge University Press), 390–418.
- Chersi, F., Mirolli, M., Pezzulo, G., and Baldassarre, G. (2013). A spiking neuron model of the cortico-basal ganglia circuits for goal-directed and habitual action learning. *Neural Netw.* 41, 212–224. doi: 10.1016/j.neunet.2012.11.009
- Chevalier, G., and Deniau, J. (1990). Disinhibition as a basic process in the expression of striatal functions. *Trends Neurosci.* 13, 277–281. doi: 10.1016/0166-2236(90)90109-N
- Cohen, J., McClure, S., and Yuo, A. (2007). Should i stay or should i go? How the human brain manages the trade-off between exploitation and exploration. *Philos. Trans. R. Soc. B Biol. Sci.* 362, 933–942. doi: 10.1098/rstb.2007.2098
- Cohen, M. X., and Frank, M. J. (2009). Neurocomputational models of the basal ganglia in learning, memory, and choice. *Behav. Brain Res.* 199, 141–156. doi: 10.1016/j.bbr.2008.09.029
- Daw, N. D., Niv, Y., and Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat. Neurosci.* 8, 1704–1711. doi: 10.1038/nn1560
- Daw, N. D., O’Doherty, J. P., Dayan, P., Seymour, B., and Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature* 441, 876–879. doi: 10.1038/nature04766
- Dearden, R., Friedman, N., and Russell, S. (1998). “Bayesian q-learning,” in *Proceedings of the Fifteenth National Conference on Artificial Intelligence (AAAI)* (Madison, WI), 761–768.
- Dimitrakakis, C. (2006). “Nearly optimal exploration-exploitation decision thresholds,” in *Proceedings of the Sixteenth International Conference on Artificial Neural Networks (ICANN 2006), Part I, Athens, Greece*, eds S. Kollias, A. Stafylopatis, W. Duch, and E. Oja (Berlin; Heidelberg: Springer), 850–859.
- Georgopoulos, A., Kalaska, J., Caminiti, R., and Massey, J. (1982). On the relations between the direction of two-dimensional arm movements and cell discharge in primate motor cortex. *J. Neurosci.* 2, 1527–1537.
- Gibson, J. (1977). “The theory of affordances,” in *Perceiving, Acting, and Knowing: Toward an Ecological Psychology*, eds R. Shaw and J. Bransford (Hillsdale, NJ: Lawrence Erlbaum and Associates), 67–82.
- Gibson, J. (1986). *The Ecological Approach to Visual Perception*. Hillsdale, NJ: Lawrence Erlbaum and Associates.
- Gold, J. I., and Shadlen, M. N. (2007). The neural basis of decision making. *Ann. Rev. Neurosci.* 30, 535–574. doi: 10.1146/annurev.neuro.29.051605.113038
- Gottlieb, J., Lopes, P. O. M., and Baranes, A. (2013). Information-seeking, curiosity, and attention: computational and neural mechanisms. *Trends Cogn. Sci.* 17, 585–593. doi: 10.1016/j.tics.2013.09.001
- Gullapalli, V. (1990). A stochastic reinforcement learning algorithm fo learning real-valued functions. *Neural Netw.* 3, 671–692. doi: 10.1016/0893-6080(90)90056-Q
- Gurney, K., Lepora, N., Shah, A., Koene, A., and Redgrave, P. (2013). “Action discovery and intrinsic motivation: a biologically constrained formalism,” in *Intrinsically Motivated Learning in Natural and Artificial Systems*, Chapter 7, eds G. Baldassarre and M. Mirolli (Berlin; Heidelberg: Springer-Verlag), 151–184. doi: 10.1007/978-3-642-32375-1-7
- Gurney, K., Prescott, T., and Redgrave, R. (2001a). A computational model of action selection in the basal ganglia. I. A new functional anatomy. *Biol. Cybern.* 84, 401–410. doi: 10.1007/PL00007984
- Gurney, K., Redgrave, R., and Prescott, T. (2001b). A computational model of action selection in the basal ganglia. II. Analysis and simulation of behaviour. *Biol. Cybern.* 84, 411–423. doi: 10.1007/PL00007985
- Gurney, K., Prescott, T. J., Wickens, J. R., and Redgrave, P. (2004). Computational models of the basal ganglia: from robots to membranes. *Trends Neurosci.* 27, 453–459. doi: 10.1016/j.tins.2004.06.003
- Harris, C., and Wolpert, D. (1998). Signal-dependent noise determines motor planning. *Nature* 394, 780–784. doi: 10.1038/29528
- Houk, J. C., Adams, J. L., and Barto, A. G. (1995). “A model of how the basal ganglia generate and use neural signals that predict reinforcement,” in *Models of Information Processing in the Basal Ganglia*, Chapter 13, eds J. C. Houk, J. L. Davis, and D. G. Beiser (Cambridge, MA: MIT Press), 249–270.
- Huk, A., and Shadlen, M. N. (2005). Neural activity in macaque parietal cortex reflects temporal integration of visual motion signals during perceptual decision making. *J. Neurosci.* 25, 10420–10436. doi: 10.1523/JNEUROSCI.4684-04.2005
- Humphries, M., and Gurney, K. (2002). The role of intra-thalamic and thalamocortical circuits in action selection. *Network* 13, 131–156. doi: 10.1088/0954-898X/13/1/305
- Joel, D., Niv, Y., and Ruppin, E. (2002). Actor-critic models of the basal ganglia: new anatomical and computational perspectives. *Neural Netw.* 15, 535–547. doi: 10.1016/S0893-6080(02)00047-3
- Kawato, M. (1990). “Feedback-error-learning neural network for supervised motor learning,” in *Advanced Neural Computers*, ed R. Eckmiller (North-Holland: Elsevier), 365–372.
- Koos, T., and Tepper, J. (1999). Inhibitory control of neostriatal projection neurons by GABAergic interneurons. *Nat. Neurosci.* 2, 467–472. doi: 10.1038/8138
- Lee, T., and Genovese, E. (1989). Distribution of practice in motor skill acquisition: different effects for discrete and continuous tasks. *Res. Q. Exerc. Sport* 60, 59–65. doi: 10.1080/02701367.1989.10607414
- Lepora, N., Fox, C., Evans, M., Diamond, M., Gurney, K., and Prescott, T. (2012). Optimal decision-making in mammals: insights from a robot study of rodent texture discrimination. *J. R. Soc. Interface* 9, 1517–1528. doi: 10.1098/rsif.2011.0750
- Marsland, S. (2009). Using habituation in machine learning. *Neurobiol. Learn. Mem.* 92, 260–266. doi: 10.1016/j.nlm.2008.05.014
- Mink, J. W. (1996). The basal ganglia: focused selection and inhibition of competing motor programs. *Prog. Neurobiol.* 50, 381–425. doi: 10.1016/S0301-0082(96)00042-1
- Minsky, M. (1961). “Steps toward artificial intelligence,” in *Proceedings IRE 49 1*, 8–30. (Reprinted in Computers and Thought, McGraw-Hill, 1963. Mapped out future research in AI with emphasis on symbolic descriptions). doi: 10.1109/JRPROC.1961.287775
- Mirolli, M., Baldassarre, G., and Santucci, V. (2013). Phasic dopamine as a prediction error of intrinsic and extrinsic reinforcement driving both action acquisition and reward maximization: a simulated robotic study. *Neural Netw.* 39, 40–51. doi: 10.1016/j.neunet.2012.12.012
- Oudeyer, P., and Kaplan, F. (2007). What is intrinsic motivation? A topology of computational approaches. *Front. Neurorobot.* 1:6. doi: 10.3389/neuro.12.006.2007
- Platt, M., and Glimcher, P. (1999). Neural correlates of decision variables in parietal cortex. *Nature* 400, 233–238. doi: 10.1038/22268
- Prescott, T., Gonzalez, F., Gurney, K., Humphries, M., and Redgrave, R. (2006). A robot model of the basal ganglia: behavior and intrinsic processing. *Neural Netw.* 19, 31–61. doi: 10.1016/j.neunet.2005.06.049
- Redgrave, P., and Gurney, K. (2006). The short-latency dopamine signal: a role in discovering novel actions? *Nat. Rev. Neurosci.* 7, 967–975. doi: 10.1038/nrn2022
- Redgrave, P., Gurney, K., and Reynolds, J. (2008). What is reinforced by phasic dopamine signals? *Brain Res. Rev.* 58, 322–339. doi: 10.1016/j.brainresrev.2007.10.007
- Redgrave, P., Gurney, K., Stafford, T., Thirkettle, M., and Lewis, J. (2013). “The role of the basal ganglia in discovering novel actions,” in *Intrinsically Motivated Learning in Natural and Artificial Systems*, Chapter 6, eds G. Baldassarre and M.

- Mirolli (Berlin; Heidelberg: Springer-Verlag), 129–150. doi: 10.1007/978-3-642-32375-1_6
- Redgrave, P., Vautrelle, N., and Reynolds, J. (2011). Functional properties of the basal ganglia's re-entrant loop architecture: selection and reinforcement. *Neuroscience* 198, 138–151. doi: 10.1016/j.neuroscience.2011.07.060
- Rosenstein, M., and Barto, A. (2004). “Supervised actor-critic reinforcement learning,” in *Handbook of Learning and Approximate Dynamic Programming, IEEE Press Series on Computational Intelligence*, Chapter 14, eds J. Si, A. Barto, W. Powell, and D. Wunsch (Piscataway, NJ: Wiley-IEEE Press), 359–380.
- Schultz, W., Dayan, P., and Montague, P. R. (1997). A neural substrate of prediction and reward. *Science* 275, 1593–1599. doi: 10.1126/science.275.5306.1593
- Shah, A., and Barto, A. G. (2009). Effect on movement selection of an evolving sensory representation: a multiple controller model of skill acquisition. *Brain Res.* 1299, 55–73. doi: 10.1016/j.brainres.2009.07.006
- Shah, A., and Gurney, K. (2011). “Dopamine-mediated action discovery promotes optimal behaviour ‘for free,’” in *Twentieth Annual Computational Neuroscience Meeting*, Stockholm. (Poster presentation. Abstract also published in *BMC Neuroscience* 2011, 12 (Suppl. 1):P138).
- Simsek, O., and Barto, A. (2006). “An intrinsic reward mechanism for efficient exploration,” in *Proceedings of the Twenty-Third International Conference on Machine Learning (ICML-06)*, eds W. Cohen and A. Moore (Pittsburgh, PA: ACM International Conference Proceedings Series), 833–840.
- Skinner, B. (1938). *The Behavior of Organisms*. New York, NY: Appleton-Century-Crofts.
- Stafford, T., Thirkettle, M., Walton, T., Vautrelle, N., Hetherington, L., Port, M., et al. (2012). A novel task for the investigation of action acquisition. *PLoS ONE* 7:e37749. doi: 10.1371/journal.pone.0037749
- Stafford, T., Walton, T., Hetherington, L., Thirkettle, M., Gurney, K., and Redgrave, P. (2013). “A novel behavioural task for researching intrinsic motivations,” in *Intrinsically Motivated Learning in Natural and Artificial Systems*, Chapter 15, eds G. Baldassarre and M. Mirolli (Berlin; Heidelberg: Springer-Verlag), 395–410.
- Stocco, A., Fum, D., and Napoli, A. (2009). Dissociable processes underlying decisions in the Iowa Gambling Task: a new integrative framework. *Behav. Brain Funct.* 5, 1. doi: 10.1186/1744-9081-5-1
- Sutton, R. (1984). *Temporal Credit Assignment in Reinforcement Learning*. Phd, Department of Computer Science, University of Massachusetts Amherst.
- Sutton, R. (1988). Learning to predict by methods of temporal differences. *Mach. Learn.* 3, 9–44. doi: 10.1007/BF00115009
- Sutton, R. S., and Barto, A. G. (1998). *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press.
- Thirkettle, M., Walton, T., Redgrave, P., Gurney, K., and Stafford, T. (2013a). No learning where to go without first knowing where you're coming from : action discovery is trajectory, not endpoint based. *Front. Psychol.* 4:638. doi: 10.3389/fpsyg.2013.00638
- Thirkettle, M., Walton, T., Shah, A., Gurney, K., Redgrave, P., and Stafford, T. (2013b). The path to learning: action acquisition is impaired when visual reinforcement signals must first access cortex. *Behav. Brain Res.* 243, 267–272. doi: 10.1016/j.bbr.2013.01.023
- Thorndike, E. (1911). *Animal Intelligence: Experimental Studies*. New York, NY: Macmillan. doi: 10.5962/bhl.title.55072
- Wickens, J. R. (2009). Synaptic plasticity in the basal ganglia. *Behav. Brain Res.* 199, 119–128. doi: 10.1016/j.bbr.2008.10.030

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 17 May 2013; accepted: 23 January 2014; published online: 11 February 2014.

*Citation: Shah A and Gurney KN (2014) Emergent structured transition from variation to repetition in a biologically-plausible model of learning in basal ganglia. *Front. Psychol.* 5:91. doi: 10.3389/fpsyg.2014.00091*

*This article was submitted to Cognitive Science, a section of the journal *Frontiers in Psychology*.*

Copyright © 2014 Shah and Gurney. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Modeling effects of intrinsic and extrinsic rewards on the competition between striatal learning systems

Joschka Boedecker^{†*}, Thomas Lampe[†] and Martin Riedmiller

Machine Learning Lab, Department of Computer Science, Faculty of Engineering, Albert-Ludwigs-University Freiburg, Freiburg, Germany

Edited by:

Marco Mirolli, Istituto di Scienze e Tecnologie della Cognizione, Italy

Reviewed by:

Mehdi Khamassi, Centre National de la Recherche Scientifique, France

Kevin Gurney, University of Sheffield, UK

***Correspondence:**

Joschka Boedecker, Machine Learning Lab, Department of Computer Science, Faculty of Engineering, Albert-Ludwigs-University, Georges-Köhler-Allee 79, D-79110 Freiburg, Germany
e-mail: jboedeck@informatik.uni-freiburg.de

[†]These authors have contributed equally to this work.

A common assumption in psychology, economics, and other fields holds that higher performance will result if extrinsic rewards (such as money) are offered as an incentive. While this principle seems to work well for tasks that require the execution of the same sequence of steps over and over, with little uncertainty about the process, in other cases, especially where creative problem solving is required due to the difficulty in finding the optimal sequence of actions, external rewards can actually be detrimental to task performance. Furthermore, they have the potential to undermine intrinsic motivation to do an otherwise interesting activity. In this work, we extend a computational model of the dorsomedial and dorsolateral striatal reinforcement learning systems to account for the effects of extrinsic and intrinsic rewards. The model assumes that the brain employs both a goal-directed and a habitual learning system, and competition between both is based on the trade-off between the cost of the reasoning process and value of information. The goal-directed system elicits internal rewards when its models of the environment improve, while the habitual system, being model-free, does not. Our results account for the phenomena that initial extrinsic reward leads to reduced activity after extinction compared to the case without any initial extrinsic rewards, and that performance in complex task settings drops when higher external rewards are promised. We also test the hypothesis that external rewards bias the competition in favor of the computationally efficient, but cruder and less flexible habitual system, which can negatively influence intrinsic motivation and task performance in the class of tasks we consider.

Keywords: striatal models, reinforcement learning model, model-free vs. model-based learning, intrinsic motivation, extrinsic motivation

1. INTRODUCTION

What motivates intelligent beings to perform certain actions in their environment is a central question in psychology. The influential paradigm of operant conditioning by Skinner (1953) held that all behavior is stimulated by external rewards presented to an animal. This view was challenged, however, by observations made by White (1959) that some behaviors are *intrinsically motivated*, i.e., they are performed simply because the activity is *intrinsically rewarding*. Deci (1971) then examined what effects external rewards would have on intrinsic motivation and found that under certain circumstances, extrinsic rewards could undermine intrinsic motivation. Later on, several studies (see extensive meta-analytic review by Deci et al., 1999) observed that external rewards can decrease cognitive flexibility in problem solving (McGraw and McCullers, 1979), and have the potential to decrease performance on complex tasks (Erez et al., 1990). These findings significantly contradicted predictions of earlier theories such as operant conditioning or utility theory in economics.

To explain these observations, several theoretical accounts have been put forward [e.g., Cognitive Evaluation Theory by Deci and Ryan (1985), Attribution Theory by Lepper et al. (1973), or Self-Determination Theory by Ryan and Deci. (2000) amongst others] which suggest different cognitive mechanisms to account for the data. However, it is not clear what *computational*

mechanisms in the brain could give rise to these phenomena. A computational model would enable quantitative comparisons of different hypotheses, test various experimental settings, and generate predictions for new, untested scenarios.

Here, we provide such a computational model by extending two previously presented models explaining behavioral control in the decision systems (Daw et al., 2005), and trade-offs between habitual and goal-directed brain processes (Keramati et al., 2011). Both of these models follow a hypothesis from behavioral economics, suggesting that two distinct control systems in the brain compete for control of actions (see e.g., Kahneman and Frederick, 2002). The models are formalized using the framework of reinforcement learning (RL, see e.g., Sutton and Barto, 1998), and it is assumed that one controller uses computationally efficient model-free RL, whereas the other one uses statistically efficient model-based RL algorithms. The model-free system represents a habitual process, implementing a cache of efficient actions for a given situation, while the model-based system realizes a goal-directed process by searching a tree of recorded state-action transition probabilities for alternative choices. Both computational models could account for several phenomena from animal experiments designed to test devaluation resistance, including habituation after extensive training, non-habituation in ambivalent tasks, and habituation in preference tasks. Our proposed

model is a mixture of both earlier models (see below for details), and, for the first time, connects them to intrinsic rewards for the model-based goal-directed subsystem. With this extension, we aim to explain three additional phenomena which the previous models could not account for.

1.1. ACTIVITY WITHOUT EXTRINSIC REWARD

When dealing with a creative or complex system, both humans and animals can be observed to interact (to “play”) with it even if no extrinsic reward whatsoever is being provided or promised.

1.2. REDUCED POST-EXTINCTION ACTIVITY

In creative tasks, the presence of strong extrinsic rewards can lead to diminished activity after said rewards have been devalued. More specifically, the activity will be lower than it would have been had the subject never received any extrinsic reward in the first place (Deci, 1971). Strong extrinsic rewards are therefore expected to suppress intrinsic motivation.

1.3. EFFECTS OF PROMISED EXTERNAL REWARDS

It has been observed that the promise of strong extrinsic rewards for a certain level of task performance does not only lead to diminished activity during creative problem solving as described above, but in fact also leads to inferior final performance on tasks involving cognitive skills (Ariely et al., 2009).

2. MATERIALS AND METHODS

Since our model is an extension of the work by Daw et al. (2005) and Keramati et al. (2011) on striatal competition, we first give a brief description of their respective approaches. After that, we will detail the changes that were newly introduced in detail.

2.1. STRIATAL COMPETITION

Both previous models intend to give a formal account of the decision system and its division into a goal-directed and a habitual module. The former realizes a model-based “tree” system that gradually builds a comprehensive model of the task, which can then be used to find an optimal sequence of steps that results in the greatest reward for a given task. In contrast, the habitual system learns in a model-free fashion as a “cache,” retaining only the knowledge of which possible action in a given situation promises a higher final payoff, but does not record which subsequent state the action would lead to. This makes it computationally cheaper than the goal-directed system, but also less adaptive to changes in the environment.

In Daw et al. (2005), these systems are assumed to be located in the prefrontal cortex and the dorsolateral striatum, respectively. While newer studies have placed the goal-directed system in the dorsomedial striatum (Yin et al., 2005), the functional distinction between the two types of system remains unchallenged.

Reproducing these aspects in the models allows them to explain several observations regarding habituation in animals. Specifically, it was found by Killcross and Coutureau (2003) as well as Holland (2004) that if a rat performs a simple lever-pulling task long enough that generates a food reward, it will become resistant to devaluation. Even if the food reward is being negated (via poison), the animal will continue performing the

same sequence of actions. If the devaluation occurs after only moderate training, no such resistance occurs, and the rat will immediately adapt its behavior.

It was argued that the observed effects are caused by the competition between both modules. The adaptable goal-directed system is active initially, but replaced by the habitual system after extended training, at which point the agent becomes resistant to devaluation. The main difference between the two models lies in the specific competition mechanism used to arbitrate between both systems.

2.1.1. Uncertainty-based competition

In the earlier model by Daw et al. (2005), it is assumed that the system is chosen which is more certain about the action to be taken. To determine uncertainty, both the model-based and the model-free system are implemented using Bayesian Reinforcement Learning Dearden et al., 1998, Mannor et al., 2004. Therefore, rather than learning Q-values for a given state, they assume a prior (Beta) distribution over Q-values for each entry in the Q-table. Bayesian updates are then used to calculate the posterior distribution based on the experience during learning. Likewise, the transition function and the terminal reward function employed by the model-based subsystem are also tables of distributions. A policy is then generated through tree-search on this model, which is realized by performing Value Iteration on a Q-function initialized to the reward function.

When the system enters some state s , the value distribution $Q_{s,a}$ is determined for each available action a . For each a , either the goal-directed or the habitual system’s estimate of $Q_{s,a}$ is used. The system that provides the Q-distribution is chosen depending on which one has the lower variance σ^2 :

$$Q_{s,a}^* = \begin{cases} Q_{s,a}^{\text{tree}} & \text{if } (\sigma_{s,a}^2)^{\text{tree}} < (\sigma_{s,a}^2)^{\text{cache}} \\ Q_{s,a}^{\text{cache}} & \text{otherwise} \end{cases} \quad (1)$$

After selecting the more confident system for each action, the actual action to be performed is chosen through Boltzmann exploration over the Q-distributions’ means μ^* , parameterized by the softmax parameter β .

$$P(a = a_i | s) \propto e^{\beta \mu_{s,a}^*} \quad (2)$$

At each time step, all distribution parameters decay exponentially with a forgetting factor θ to their priors, thus keeping the system capable of learning from new experiences even after long training durations.

Since the tree-search is performed until convergence at each time step, a sudden change in the reward model resulting from a devaluation event will immediately be propagated all the way through the state space. In contrast, the model-free system will have to perform the original sequence several times to register a change in the terminal state’s value in the starting state. The habitual system becomes dominant after extended training, but not after moderate one, since its variance decreases more slowly than that of the goal-directed system. Thus, the model accounts for the empirical findings.

2.1.2. Value-based competition

Keramati et al. (2011) modify the basic approach of Daw et al. (2005) by using the value of perfect information (VPI) instead of uncertainty. Here, the model-free system computes how much value would be gained from knowing the true value of a given action. Such knowledge would only have value if it allows the agent to improve its policy. Therefore, it should reveal that the previously preferred action is not in fact optimal, either by showing that its true value is less than thought, or that another action promises higher rewards. Formally, the gain G of knowing that an action a has the value $Q_{s,a} = x$ can be computed as follows, where the calculation differs depending on whether a is the optimal action a_1 or second best action a_2 as judged by the habitual system thus far.

$$G_{s,a}(x) = \begin{cases} Q_{s,a_2}^{\text{cache}} - x & \text{if } a = a_1 \text{ and} \\ & x < Q_{s,a_2}^{\text{cache}} \\ x - Q_{s,a_1}^{\text{cache}} & \text{if } a \neq a_1 \text{ and} \\ & x > Q_{s,a_1}^{\text{cache}} \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

The VPI is then simply given by the expected Gain over the distribution of possible values that $Q_{s,a}$ can take.

$$\text{VPI}(s,a) = E[G_{s,a}(x)] \quad (4)$$

Intuitively and generally speaking, this value is higher if an action's Q -distribution overlaps strongly with the best action, since in this case the former may turn out to be preferable. Conversely, once the distributions have separated, knowing the true value of an action is unlikely to change which one is ultimately chosen.

Once computed, the VPI is compared against the costs of opportunity for performing a tree-search, denoted by $\bar{R}\tau$, with \bar{R} being the expected average reward and τ being the cost in terms of deliberation time for traversing an edge of the tree.

$$Q_{s,a}^* = \begin{cases} Q_{s,a}^{\text{tree}} & \text{if } \text{VPI}(s,a) > \bar{R}\tau \\ Q_{s,a}^{\text{cache}} & \text{otherwise} \end{cases} \quad (5)$$

Only if the VPI is higher than the opportunity costs is the model-based system activated to determine the true reward. The winning system's estimate is then used for action selection. Since determining the VPI does not involve the goal-directed system in any way, this approach better adheres to the assumption that using the habitual system is less time-intensive.

Finally, the average reward \bar{R} is updated with new observations using learning rate η :

$$\bar{R}_{t+1} = (1 - \eta)\bar{R}_t + \eta r_t \quad (6)$$

One advantage of using the VPI instead of both modules' uncertainty lies primarily in considerations of speed. Since the VPI can be computed purely from the habitual system's uncertainty about the value distribution, thus often eliminating the need

for the costly computations required when activating the goal-directed system. In contrast, the previous model always required the calculation of the goal-directed system's uncertainty and value. Without the ability to speed up the decision process, that would raise the issue of why a habitual module should even have evolved.

It is worth noting that the goal-directed system used both here and by Keramati et al. (2011) does not initially provide perfect value estimates, making the term "value of perfect information" somewhat incorrect. As such, it may not fulfill its purpose of improving the action choices at the very beginning of the learning process. However, its ability to reason globally allows it to learn sensible actions from fewer observations than the rigid cache, and thus to provide value estimates soon.

2.2. MODEL EXTENSION

Aside from using a mixture of the features present in our predecessor models, there are two major extensions in our model that were not present in its predecessors, which will be described in detail in the following.

2.2.1. Intrinsic rewards

The main contribution of our model lies in its extension with a mechanism for intrinsic motivation. The central feature of intrinsic rewards lies in that their value depends on the current state of the model, as opposed to extrinsic rewards that are provided by the process or environment. As such, intrinsic rewards can notably arise *only* in the goal-directed system, and are not applied to the habitual one.

Currently we consider only one of multiple types of intrinsic reward, namely the learning progress of the transition model (similar to Oudeyer et al., 2007). Learning progress is based on the intuition that a system should explore regions where it can currently learn the most based on the state of its internal models, i.e., make the largest progress at improving its models. In contrast, simple metrics based on surprise are prone to get stuck in completely unpredictable situations which is avoided by rewarding progress (i.e., reduction of surprise over time) instead. There are other proposed aspects to intrinsic motivation, such as competence-based and information-theoretical mechanisms (for an overview, see section 4.1), but we focus on progress for the sake of simplicity, as it already accounts for the phenomena we consider by itself. As measure of learning progress we use the magnitude of shifts in the means of the transition function's distributions. Formally, the intrinsic reward I for choosing action a in state s is given by the equation:

$$I_{s,a} = \iota \sum_{s' \in S} |\Delta \mu_{s,a,s'}^{\text{trans}}| \quad (7)$$

Here, ι is a factor used to accentuate the intrinsic rewards and bring them into the same order of magnitude as the extrinsic ones.

I is then added to the result of the tree-search:

$$\tilde{Q}_{s,a}^{\text{tree}} := Q_{s,a}^{\text{tree}} + I_{s,a} \quad (8)$$

The resulting Q -values \tilde{Q}^{tree} are then used in place of those determined by the search for the purpose of subsystem selection and exploration.

At this point, one may wonder why, from among the many alternative types of intrinsic motivation, we choose $\Delta\mu^{\text{trans}}$ rather than $\Delta(\sigma^2)^{\text{trans}}$, which provides a more meaningful measure of learning progress. Our model provides the variance readily, but when using Dirichlet distributions with a large number of states, the variance is not a useful metric. This is because shifts in variance for observations that have or have not been made before differ very little. Only once the transition model is nearly stable will unexpected observations cause a distinct shift. However, by that time, the intrinsic rewards will be too low to have significant influence on action selection anyway.

2.2.2. Transition costs

Aside from intrinsic rewards, we also introduce transition costs. While a common element of RL and formalized in the Bellman Equation (see Sutton and Barto, 1998), they were not present in the model by Daw et al. (2005). Instead, the entire terminal reward of a trajectory was propagated all the way to the starting state.

By accommodating them, we enable the model to acquire minimum-time policies in tasks where trajectories can contain loops. Most importantly, transition costs can also be chosen differently for each action, thereby modeling energy conservation.

It is worth noting that action-based transition costs do not fall cleanly into the distinction between extrinsic and intrinsic rewards. Traditionally considered extrinsic rewards, they are likewise applied to the habitual system, as opposed to intrinsic rewards, which due to being model-based can naturally only occur within the goal-directed system. On the other hand, they mimic intrinsic rewards in that they are essentially inherent—one may be tempted to say “intrinsic”—to the agent. Action costs are not provided by the environment, and can thus be assumed to occur even when other extrinsic rewards do not. To avoid confusion, we will dub them *action rewards* in the following and mention explicitly when they are used and when not, since their appearance is not bound to either of the two major reward types.

Applying transition costs can easily be done during both tree-search and update of the habitual system by adding them to the discounted extrinsic reward that would result from choosing the optimal action a_* in the successor state s' . Doing so yields a new target mean $\hat{\mu}$:

$$\hat{\mu}_{s, a} = \gamma\mu_{s', a_*} + r_a \quad (9)$$

The update rule for the distribution parameters also requires the second moments of the successor states' Beta distributions. We therefore generate a new distribution $\hat{Q}_{s, a} = \text{Beta}(\hat{\alpha}, \hat{\beta})$ with the target mean $\hat{\mu}_{s, a}$, from which we can then infer these moments. Between its parameters, the following relationship must hold:

$$\frac{\hat{\alpha}}{\hat{\mu}} = \frac{\hat{\beta}}{1 - \hat{\mu}} \quad (10)$$

Thus, we need to fix one of the Beta parameters to determine the other. Depending on which one is chosen, the distribution's variance may either increase or decrease, as illustrated in **Figure 1**. Under the reasonable assumption that every step of tree-search introduces additional uncertainty, we choose whichever would cause a variance increase.

The resulting $\hat{Q}_{s, a}$ is then used for the computation of the new distribution parameters. Analogously to Daw et al. (2005), they are updated using a mixture rule derived from Dearden et al. (1998).

$$\int_0^1 \text{Beta}(\alpha_{s, a} + x, \beta_{s, a} + (1 - x)) \hat{Q}_{s, a}(x) dx \quad (11)$$

Details on the closed-form update can be found in the supplemental material to Daw et al. (2005).

2.2.3. Model mixture

Like Keramati et al. (2011), we use the VPI to mediate between the goal-directed and the habitual subsystem. The alternative approach of using the variance of the Q -function's estimates would not be plausible in a framework containing intrinsic rewards. Intrinsic motivation is generally assumed to be high

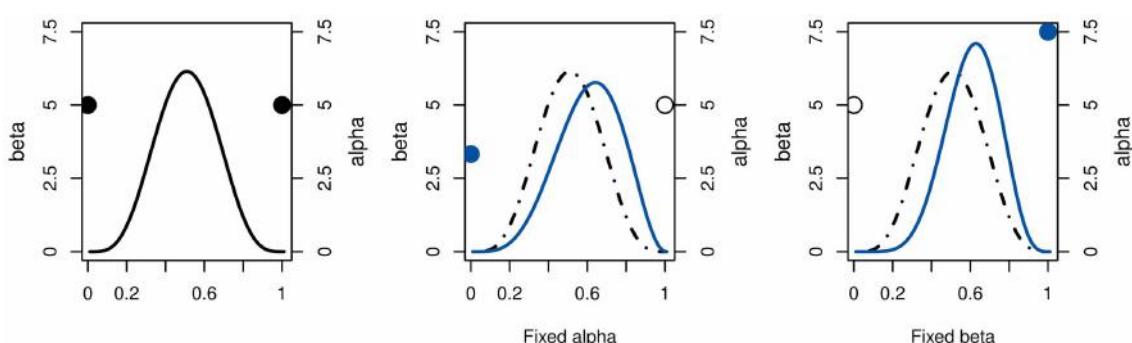


FIGURE 1 | Illustration of the relationship between Beta parameters for a given positive mean shift of an arbitrary distribution (left). If α is fixed (middle), the lower β results in a flatter distribution

with higher variance. Conversely, fixing β would reduce the variance (right). For action costs, i.e., negative action rewards, the effects are reversed.

for regions of the state space in which the model has not been learned yet. In these regions, the goal-directed system's variance will also be particularly high (Oudeyer and Kaplan, 2007). If the goal-directed system's variance is involved in the competition mechanism, this will lead to it being rejected in precisely those situations when intrinsic motivation is high, thereby neutralizing the effect of the latter.

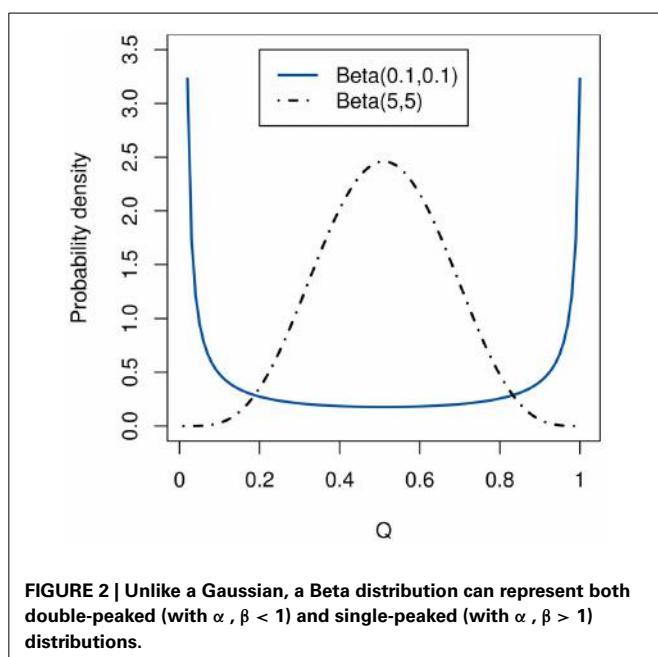
From the original approach by Daw et al. (2005) we retain the use of Beta and Dirichlet distributions to represent the model and the policies learned by the agent, as opposed to the Gaussians used by Keramati et al. (2011). Using a Beta distribution for the policy carries the advantage that its probability density function can have two peaks, as illustrated in **Figure 2**. Therefore, it is able to represent a limited amount of ambiguity arising from non-determinism, while single-peaked Gaussians model only uncertainty. Since their range is constrained in the interval [0; 1], we can simply compute the integral of the VPI by sampling.

$$E[G_{s,a}(x)] = \int_0^1 G_{s,a}(x)P(Q_{s,a}^{\text{cache}} = x) dx \quad (12)$$

Instead of Boltzmann exploration, we employ ϵ -greedy exploration when choosing an action, i.e., at each decision point, a random action is uniformly sampled from the options with probability ϵ . This approach was chosen because given more complex tasks, the different learning speed of both subsystems may cause their Q-values to be of considerably different magnitude. In such cases, Boltzmann exploration is implausible, as it would virtually eliminate the chance of attempting an underestimated action, and thus prevent the system from learning its true value.

3. RESULTS

Our model is evaluated in a number of settings, which can be divided into two broad classes. The first consists of variations of a simple feeder task, identical to those by Daw et al. (2005),



which are to show that even with the modifications introduced here, the model still reproduces the basic devaluation resistance effects of its predecessors. Afterward, we will examine our central phenomena related to intrinsic motivation and activity in a more complex, "creative" task. Here, we take creative to mean a problem that requires a long chain of actions to solve, where each action does not cause a visible approach toward the goal.

3.1. DEVALUATION RESISTANCE

Daw et al. (2005) and Keramati et al. (2011) mostly examined their respective models using a decision task inspired by experiments with rats (Holland, 2004, Killcross and Coutureau, 2003), where the animals needed to manipulate a feeding apparatus in a short sequence to generate a reward. Those sequences had a maximum length of two decision points, and either two or three possible actions were available.

The first, simpler variant of the task allows the agent to choose between two actions, representing a lever press and a magazine entry. Only a press followed by an entry generates any reward, while any other sequence leads to a restart. In a second variation of moderate difficulty, there is an additional chain-pulling action, which, if followed by a magazine entry, leads to a different, but equivalent, extrinsic reward.

We perform the same series of experiments, with largely identical parametrization. Those that were changed, as well as newly introduced ones, are summarized in **Table 1**.

To examine the system's habituation, we devalue the goal state that is reached through the lever press by resetting its extrinsic reward distribution to Beta(1, 15). This is done after both moderate (20 episodes) and extensive training (200 episodes), and the changes in the ratio at which the lever is pressed is observed. In the moderately difficult setting, the devaluation takes place slightly later after 240 episodes to account for the more difficult task.

The results for all settings, summarized in **Figure 3**, are consistent with those of the predecessor models. While the system generally reacts more quickly to an early devaluation in the simple setting, its behavior does not change readily after extensive training, due to the inflexible habitual system having become active. The effect of early devaluations exhibits a much higher variance, which is to be expected; considering the random nature of exploration, the degree to which the system has learned the optimal policy and become habituated can differ considerably after a mere 30 episodes.

The speed of adaption mirrors the rate at which the goal-directed system was used around the time of devaluation, as **Figure 4** illustrates. In the moderate task, the ambiguity of the two available actions causes a persistently high VPI and thus a continued use of the goal-directed system. Coupled with the

Table 1 | Default parameters that were used in the feeder task.

Parameter	Symbol	Value
Search costs	τ	0.1
Exploration	ϵ	0.2
Intrinsic reward factor	ι	2.0

high accuracy of the transition model after extended training, this allows the agent to switch to the chain-pulling action immediately.

3.2. REWARD-BASED ACTIVITY

The above feeding tasks consist only of very few states and actions, making them too simple to showcase those phenomena related specifically to intrinsic motivation. We therefore consider a more complex setting, adapted from the Playroom environment used

by Singh et al. (2005), albeit simplified to accommodate the use of exact inference Bayesian RL.

In this task, the agent has to learn to manipulate a number of objects, each of which causes a different effect when used. A blue box can be used to start playing music, while a red one stops it. A switch toggles the lighting of the room, which causes the colored boxes to become indistinguishable. Lastly, there is a toy monkey, which does not cause any effect and serves as a neutral distractor. These objects need to be used in a specific sequence to bring about some desired goal state, which differs between experiments. Generally, the goal is to turn the music on and the light off, with additional success requirements in some settings.

The agent possesses a hand and an eye, both of which must rest on an object for it to become usable. Aside from performing an object affordance, the agent can also move its eye to a random object, bring the hand to the object the eye is resting on, or perform a null action that has no effect whatsoever. The null action generates a small positive action reward, unlike the other actions which cause negative ones. We thereby model an agent's general tendency to prefer the action that exerts the least effort.

While still simple for a task aimed at intrinsic motivation, it is considerably more complex than the food dispensal experiments. Most notably, trajectories can be cyclic, and one of the actions is non-deterministic. In addition, the partial observability of the state when the light is off can lead to local minima in the policy.

In this framework, we observe the behavior of the system using different combinations of intrinsic and extrinsic rewards, and determine whether the phenomena described in section 1 can be reproduced. Action rewards are present in all cases.

Unless noted differently, the system was parameterized as in **Table 2**. Most notably, the forgetting factor θ , the reward horizon η and the Dirichlet initialization α_i were adjusted to account for the longer episodes and more complex process model; otherwise, the system would forget old experiences faster than it could collect new ones. The action rewards r_a were always very slightly positive (0.005) for the null action, and negative (-0.02) for all others. They thereby model the intuitive assumption that if doing nothing promises the same reward as performing an action, the null action should be preferred. At the same time, the use of the null action should not accumulate too high action rewards, lest it overshadow those arising in the terminal states, where a terminal extrinsic reward of 1 was given.

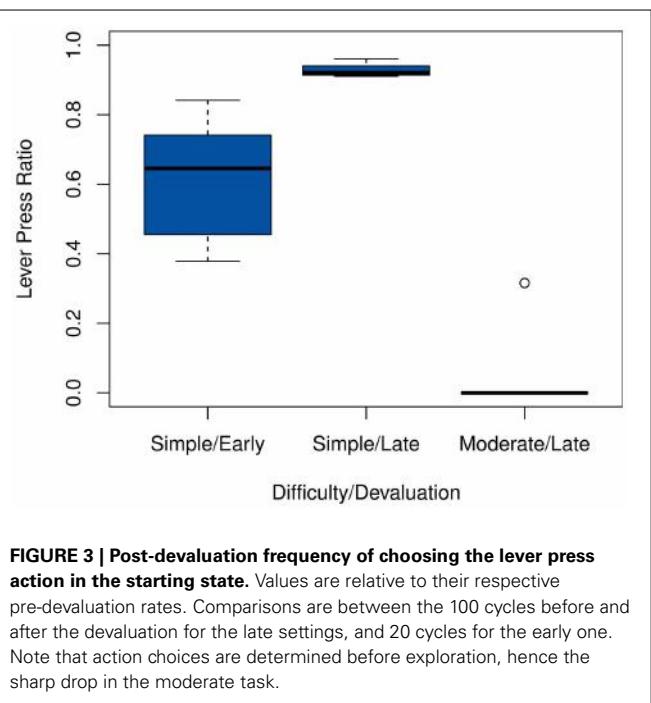


FIGURE 3 | Post-devaluation frequency of choosing the lever press action in the starting state. Values are relative to their respective pre-devaluation rates. Comparisons are between the 100 cycles before and after the devaluation for the late settings, and 20 cycles for the early one. Note that action choices are determined before exploration, hence the sharp drop in the moderate task.

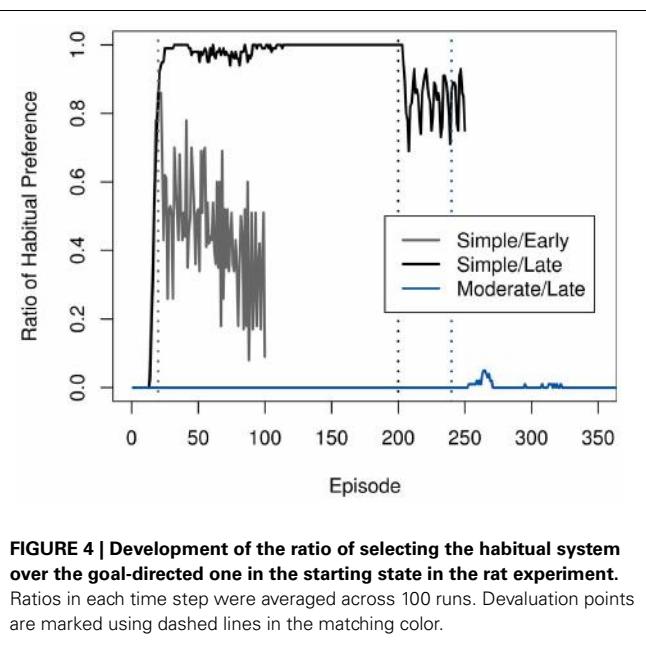


FIGURE 4 | Development of the ratio of selecting the habitual system over the goal-directed one in the starting state in the rat experiment. Ratios in each time step were averaged across 100 runs. Devaluation points are marked using dashed lines in the matching color.

Table 2 | Default parameters that were used in the Playroom task unless noted otherwise.

Parameter	Symbol	Value
Forgetting factor	θ	0.9999
Search costs	τ	0.1
Update rate of avg. reward	η	0.001
Exploration	ϵ	0.2
Intrinsic reward factor	ι	2.0
Initial transition model	α_i	0.1
Reward discount factor	γ	0.95

3.2.1. Activity without extrinsic rewards

A first experiment compares the activity of the system with and without intrinsic rewards. In this setting, there are no external rewards whatsoever, aside from the action-dependent transition costs. One would intuitively expect the overall activity, i.e., the occurrence of non-null actions, to be increased when using intrinsic rewards—higher motivation should naturally lead to more activity. And indeed, as **Figure 5** illustrates, their use leads to a significantly lower rate at which the null action is chosen.

The activity with intrinsic motivation drops to a similar level as without it only after extensive training, once the model has stabilized and no more intrinsic reward can be generated. This effect seems plausible as well, seeing as how even a motivated individual will eventually cease playing or being otherwise active. It is caused by the retaining of action-dependent costs, which will always cause the system to settle on the null action in the end.

3.2.2. Post-extinction activity

To show that stronger extrinsic rewards lead to less activity, as proposed in section 1.2, we next have the system learn a policy while providing the maximum extrinsic reward upon entering the goal state s_+ . In this case, s_+ is reached by having the music turned on and the lights off. We devalue it either after 50 or after 200 episodes of training by replacing the distribution of the extrinsic reward model for the goal state with the Beta distribution Beta(1, 15). The parameters of the replacement distribution were chosen in accordance with Daw et al. (2005) in such a way as to concentrate most of the probability mass at 0. Note that we devalue the goal, rather than merely extinguishing its extrinsic reward, under the assumption that for higher-level intelligent agents, an extinction will be registered immediately, like a devaluation.

If the devaluation occurs early, the post-devaluation activity drops sharply compared to its earlier level, as shown in **Figure 6**. In contrast, the purely intrinsic system remains active during the same time period. Only considerably later, once all intrinsic motivation in the model has been exhausted, does it become as inactive as the system using extrinsic rewards does after the devaluation.

The lowered activity is in fact caused by the re-activation of the goal-directed system. As the costs of opportunity for performing a tree-search decrease, it takes over from the habitual system as seen in **Figure 7**. The previous takeover of the habitual system caused the agent to be active mostly in a limited region of the state space, as any exploration attempts were cut short by the

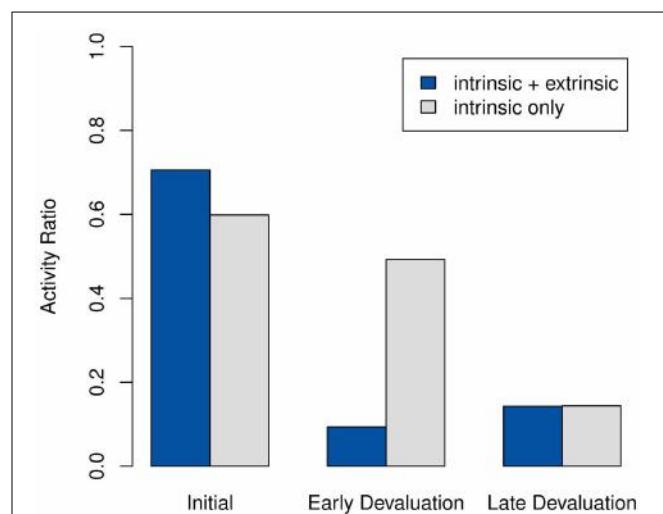


FIGURE 6 | Percentage of non-null actions chosen by the system using both intrinsic and extrinsic rewards, compared to activity using only intrinsic motivation. The goal state is devalued after episodes 50 and 200.

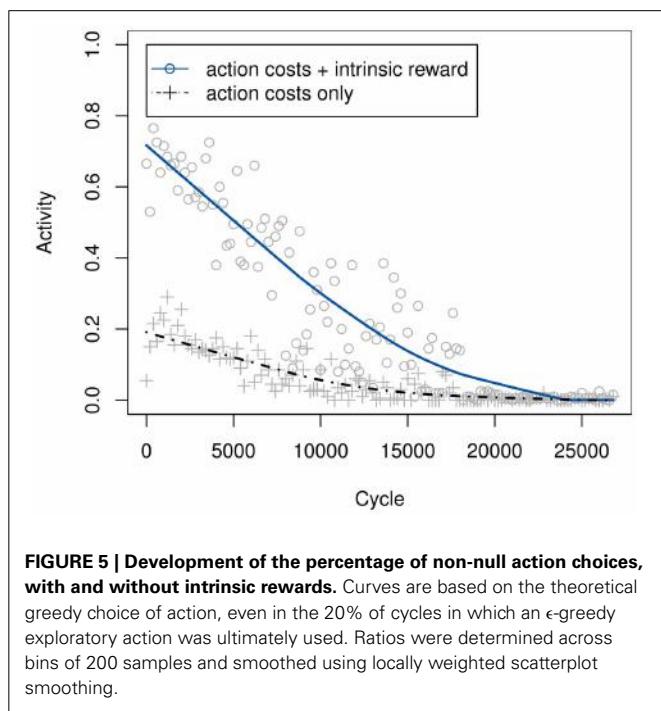


FIGURE 5 | Development of the percentage of non-null action choices, with and without intrinsic rewards. Curves are based on the theoretical greedy choice of action, even in the 20% of cycles in which an ϵ -greedy exploratory action was ultimately used. Ratios were determined across bins of 200 samples and smoothed using locally weighted scatterplot smoothing.

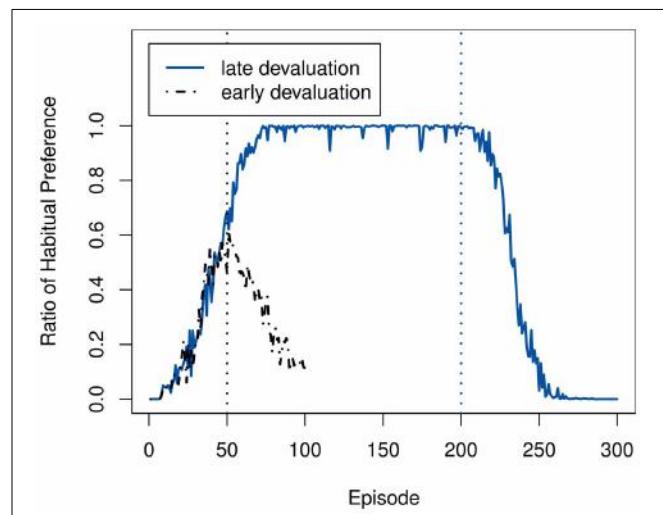


FIGURE 7 | Ratio of how often the habitual system is selected vs. the goal-directed one, when using both intrinsic and extrinsic rewards. The vertical lines mark the times of devaluation at 50 and 200 episodes.

habitual system's drive to reach the goal. Consequently, the model in this area of the state space is very accurate already. Therefore, no intrinsic reward is generated anymore, and the goal-directed system will not deviate from its path once having taken over. Essentially, due to the prolonged activation of the habitual system, the intrinsic motivation will have been exhausted without having the chance to cause any increased exploration and activity.

Also note that in **Figure 6** the purely intrinsic setting results in slightly lower initial activity than the pre-devaluation case. This observation seems plausible, since a system not driven by extrinsic rewards would be more likely to try the sub-optimal null action to improve its model.

3.2.3. Scope of motivation

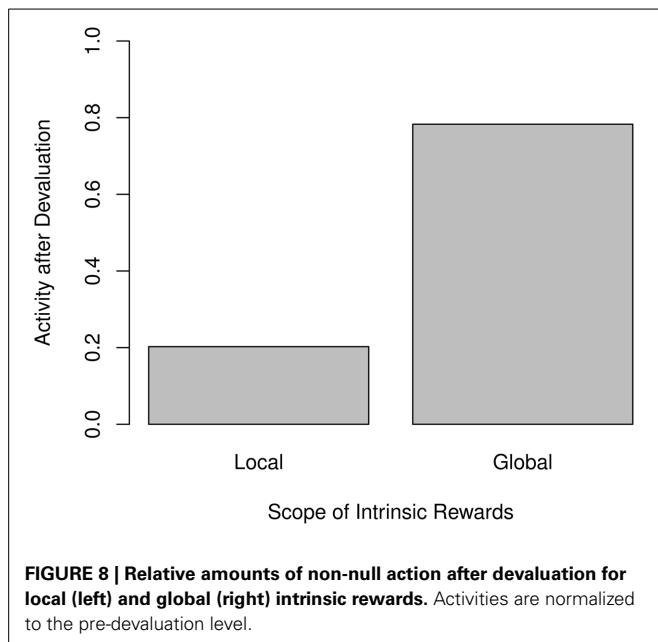
One assumption we made was the local scope of intrinsic motivation. In accordance with Equation (8), the intrinsic reward I is only applied to the final Q^{tree} after the tree-search. Therefore, only the progress in the transition model out of the current state is considered when generating I .

One possible alternative would be to not consider intrinsic rewards locally, but globally, by applying them to the target mean already during tree-search. To do so, one would merely have to revise Equation (9) to

$$\hat{\mu}_{s, a}^{\text{tree}} = \mu_{s', a_*}^{\text{tree}} + r_a + I_{s, a} \quad (13)$$

This should drive the agent more strongly into areas of the state space it has not observed yet, facilitating the acquisition of a better model.

However, the assumption of global intrinsic rewards is inconsistent with the empirical findings. **Figure 8** compares the post-devaluation activity between both approaches, and it becomes clearly apparent that the previously observed reduction in activity becomes much less pronounced when using global motivation.



3.3. SYSTEM PERFORMANCE

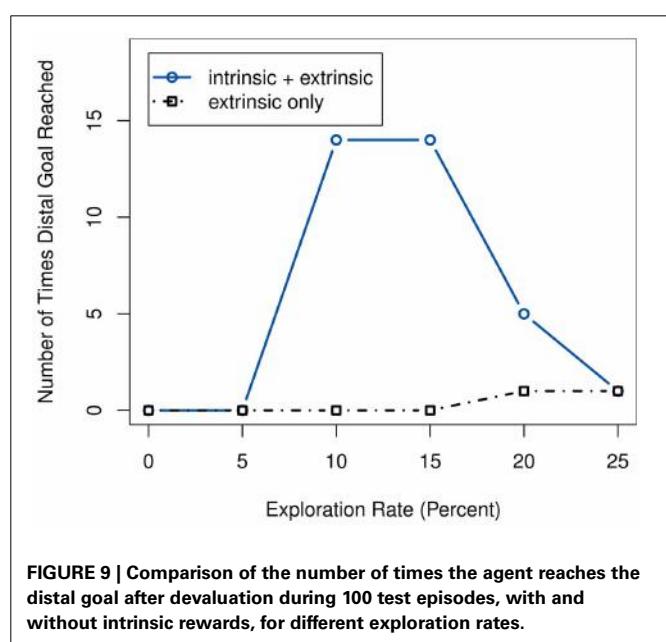
While intrinsic rewards as modeled here account for the above activity phenomena, there have been no considerations of learning performance. Therefore, we also perform a number of experiments in the same setting as before to examine the overall performance of the system.

3.3.1. Model acquisition

Clearly, a sensible model of intrinsic rewards should also justify their existence, as one would expect them to aid learning in some manner.

We thus test the system with and without intrinsic motivation in a task in which we change the goal state after a period of training. Initially, the agent receives an extrinsic reward for turning on the music and switching off the light as before, regardless of where the hand and eye are placed when the two conditions are met. After 200 episodes, one variation of the goal is devalued: if hand and eye are on a box when the music and light conditions have been met (i.e., if the light has been left off), no more extrinsic reward is given. The other possibility of having the hand and eye on the switch at the time, i.e., (turning the light on before manipulating the music, then turning it off again) remains as before. The first combination, which we will refer to as the *proximal goal*, can be potentially reached in as little as three steps, while the second *distal goal* requires three times as many.

The setup is repeated both using intrinsic rewards and using only extrinsic ones. We observe the frequency at which the agent manages to reach the remaining goal state after the devaluation. While it almost never enters the distal when only extrinsic rewards are given, it does manage to do so more often if using intrinsic motivation. The effect is not completely independent of the ϵ -greedy exploration; as figure **Figure 9** illustrates, even the intrinsically motivated system fails to find the distal goal in case of too low a value for ϵ . Similarly, excessive over-exploration causes



the performance to drop as well, as it prevents the agent from performing its learned policy. Regardless, the system clearly performs better with intrinsic reward than without, and this effect is even more pronounced if using slightly lower values for ϵ .

The results can be explained by the intrinsically motivated system's drive to better explore the state space. Thus, it possesses a higher chance of finding the distal goal state. Ideally, the agent should then directly learn to prefer the distal route, as it provides a guaranteed extrinsic reward—unlike the proximal one, due to the inability to differentiate the box colors while the light is off. But even if it does not, having seen this alternate goal would enable it to immediately switch over to it once the devaluation occurs.

3.3.2. Effects of promised rewards

As a final experiment, we examine how the model accounts for phenomena related to promises of extrinsic rewards. As observed by Ariely et al. (2009), a high expectation of being rewarded later upon completion of a task can actually reduce an agent's performance in complex tasks compared to a purely intrinsically motivated individual.

A reasonable assumption to simulate promises of later rewards seems to be to fix the average reward \bar{R} at 1, i.e., treat the promise of extrinsic rewards just the same as their observation. We thus take \bar{R} as the *expected* reward, rather than the *observed* average. In fact, this assumption is closer to those of Niv et al. (2007), whose model of tonic dopamine levels Keramati et al. (2011) have based the concept of \bar{R} on.

The conditions with and without fixed \bar{R} are compared with respect to both the time spent on reasoning processes and the ability to learn a task. We also test in two different settings of distinct difficulty, both of which require the agent to turn the music on and the light off while looking at the blue box. In the *distal* task, the agent starts in the same configuration as before, with light and music off, while in the *proximal* setting, the music is already playing and the light is on, requiring it to perform a much simpler sequence of actions.

The results for 300 episodes of training are summarized in Figure 10. Using promises of extrinsic reward reduces the amount of time spent on tree-searches significantly, particularly in the distal setting. However, the speed improvement also comes at a drastic decrease in performance in complex tasks, with the fixed \bar{R} completely preventing the agent from solving the distal case.

It should be noted that this behavior does not result from our additions to the model, but would already have been present in that of Keramati et al. (2011) using the slightly changed interpretation of \bar{R} adopted here. It is included here mostly because it has been ignored in the prior work, despite its noteworthy consistency with the empirical findings of Ariely et al. (2009).

4. DISCUSSION

We have proposed an extension to two previous models of the striatal learning system that introduces the concept of intrinsic motivation. By assigning additional intrinsic rewards for higher learning progress, we were able to reproduce several additional empirical phenomena that were not covered by our predecessors. In particular, we account for the fact that the presence of intrinsic motivation predictably raises the overall activity, but that it can be

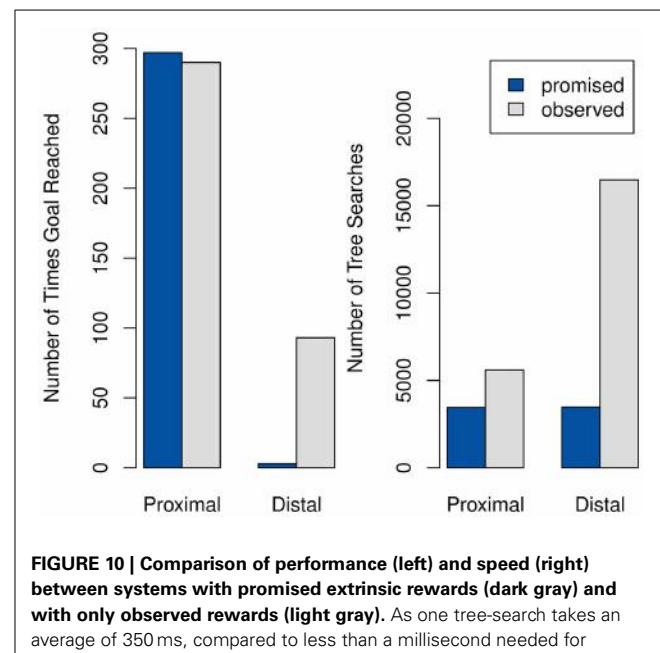


FIGURE 10 | Comparison of performance (left) and speed (right) between systems with promised extrinsic rewards (dark gray) and with only observed rewards (light gray). As one tree-search takes an average of 350 ms, compared to less than a millisecond needed for querying the habitual system, the number of tree-searches is directly proportional to the total time.

suppressed by high extrinsic rewards in turn. We have also shown that intrinsic rewards lead to better system performance in more complex tasks requiring creative solutions.

Of course, there are always aspects of the model that could be improved or require clarification, as well as behaviors that have not been examined empirically yet. These will be discussed in more detail in the following.

4.1. COMPUTATIONAL INTRINSIC MOTIVATION AND RELATED BIOLOGICAL MODELS

The principles of intrinsically motivated learning have gained increasing interest in the field of computational RL. Formalization of different aspects of intrinsic motivation, such as curiosity or competence, are expected to provide general, task-independent mechanisms that let artificial agents explore their own skills and their environment efficiently and autonomously. Furthermore, the models, which the agents build through environment interaction guided by intrinsic motivations, promise to enable improved adaptability to environmental changes or new task requirements.

Starting with the pioneering work of Schmidhuber (1991a,b), who introduced curious model-building controllers that got rewarded strongest for (near-mismatches of) predictions about the world, to Singh et al. (2005) who used internal reward signals proportional to the agent's error in predicting salient events in a related way, many approaches that tried to formalize notions of interestingness, curiosity, competence, and improvement of an agent's model about the world have been proposed [e.g., by Oudeyer et al. (2007), Schembri et al. (2007), Schmidhuber (2008), Baranes and Oudeyer (2009), and Grzyb et al. (2011)]. Here, our focus is on computational mechanisms that could explain phenomena observed in the psychology literature, i.e.,

on cognitive modeling, rather than proposing general-purpose reward mechanisms. A full overview of approaches from the computational RL literature is therefore beyond the scope of this article. Surveys for this purpose, however, such as Oudeyer and Kaplan (2007) and Schmidhuber (2010), as well as the recent book by Baldassarre and Mirolli (2013) give a much more complete picture in this regard.

Recent models dealing with aspects of intrinsic motivation from a biological perspective include those in Bolado-Gomez and Gurney (2013) and Mirolli et al. (2013). Both of these propose a role for the phasic dopamine signal from dopaminergic neurons in the brain, and both strive for consistency with neuroscientific data. In Bolado-Gomez and Gurney (2013) the authors suggest that this signal indicates surprising actions outcomes, and that objects associated with such outcomes acquire a novelty salience. They show that these signals can be used by an agent for the purpose of action discovery. In Mirolli et al. (2013), on the other hand, it is proposed that phasic dopamine signals reward prediction errors which are shaped by two different kinds of reinforcers: temporary, internal rewards for unexpected stimuli the agent experiences, and permanent, external rewards of a biological nature. Based on this assumption, phasic dopamine can drive both discovery and learning of new actions in a unified way. The model we present here also relies on an internal reinforcer which the agent can perceive in case its model of the world changes (see below). However, at this point, we do not identify the exact source of this signal. In future studies, it might be interesting to examine whether our proposed mechanism would fit the empirical data about phasic dopamine release though.

4.1.1. Alternative mechanisms of motivation

The underlying assumption behind our concept of intrinsic motivation is higher learning progress yields increased rewards. To measure progress, we observed shifts in the model's distribution means. This approach is inferior to tracking reductions in the distribution variance, as it does not allow us to differentiate between actual learning progress and cases where the model simply cannot be learned, for instance due to non-determinism. However, as described in section 2.2.1, the variance cannot be used when using Dirichlet distributions. Therefore, future improvements should try to either replace the distribution type used, or examine if alternate types of intrinsic motivation still exhibit the same behavioral effects.

4.1.2. Applicability to larger problems

In this work, we were focused purely on the explanation of empirical phenomena. For the sake of a clean theory, we used exact inference Bayesian RL. However, this approach quickly becomes intractable when applied to more complex problems. Both from a pragmatic standpoint as well as from a theoretical one—after all, rats and humans are capable of solving problems more difficult than pressing a lever or manipulating a small number of objects in sequence—it would therefore be desirable to replace it with approximative methods. Ideally, the observed phenomena should remain in that case. The ability to solve more complex tasks would also enable us to truly examine the validity of the model and of different hypotheses of motivation quantitatively.

4.1.3. Isolated treatment of actions

In our model, just like in those of our predecessors, we assume that the choice between the habitual and the goal-directed system is made independently for each available action, and only afterward exploration is performed over the resulting Q -values. Therefore, once the VPI approaches the threshold $\bar{R}\tau$, the habitual system may take over for single actions, but not for others. This can potentially lead to sub-optimal behavior if both systems learn at different speeds, as is often the case for complex tasks. If, then, the goal-directed system has a lower estimate than the habitual one, its prediction will be disregarded during exploration, despite generally being more accurate.

While this effect does not usually prevent learning, as either the sub-optimal action will also drop to its true level over time, or its VPI will decrease below threshold, this may reduce the speed at which the system learns to solve a task. Thus, for practical applications, one might either use the goal-directed system to determine all actions' values if even one calls for it, or re-calculate the VPI for all actions immediately after performing a tree-search. These approaches should still account for all observed phenomena, and may be worth examining in future works.

4.1.4. Integration with other models

A model for the division of the decision-making system in rats has also been proposed by Caluwaerts et al. (2012), albeit with the goal of explaining a different type of behavior entirely, namely navigation. While their design of a learning arbitration mechanism does not readily afford a speed/accuracy trade-off as introduced by Keramati et al. (2011), their use of learning progress to detect context changes (i.e., a shift in the goal state) could prove compatible with our model and potentially be employed to replace the explicit devaluation used so far.

4.2. PARAMETRIZATION

The model was generally designed to be robust to the choice of its parameters. Usually, their exact values should only affect the speed at which the system learns and the time at which the observed phenomena occur. However, there are a few parameters that influence the principal behavior of the system.

4.2.1. Search costs

In our model, we adopted the VPI-based competition mechanism of Keramati et al. (2011) for its high plausibility and larger compatibility with intrinsic rewards. It should, however, be noted that the choice of the active subsystem in Equation (5) depends heavily on the search costs τ . Since the habitual system's value distributions may always overlap to some extent, the VPI will generally converge to some non-zero value. Thus, if τ is chosen too small, the habitual system may never become active as the tree-search can be performed practically for free. Conversely, too high a search cost will prevent the goal-directed system from being chosen. The fact that the same setting of $\tau = 0.1$ can be used both for the simple feeder task and the more complex Playroom suggests that the admissible range of τ is wide enough to not require an exhaustive search. Even so,

in principle it may be necessary to choose τ appropriately in different tasks.

4.2.2. Forgetting factor

One aspect that the system is fairly dependent on is the forgetting factor θ . With $\theta = 0.98$, as used by Daw et al. (2005), it is impossible to learn a task as complex as the Playroom setting, since the distribution parameters will usually decay back to their priors faster than new experiences are acquired. This requires us to tune the parameter closely to the task complexity.

In this work, we settled for a setting of $\theta = 0.9999$, therefore practically turning parameter decay off. This approach comes at a cost, in turn, as it makes it difficult to change the system's behavior after a while. Once the distributions have stabilized after extensive training, new experiences will be virtually ignored. Also, when learning tasks with a larger state space, the acquisition of the Dirichlet transition model may take noticeably longer than learning a policy along a narrow trajectory in the habitual system, causing the latter to become severely over-trained.

While such behavior can actually be realistic—after all, a habit usually takes very long to unlearn—it would effectively render the habitual system useless in real-world applications. For its existence to be truly plausible, the system needs to be extended with a more robust mechanism for forgetting experiences. One option would be a surprise-based approach, which causes the parameter decay to accelerate when an unexpected event occurs, while gradually slowing down otherwise.

4.3. PREDICTIONS

Our model makes a number of assumptions and shows behaviors that have not been examined in empirical studies to date. These predictions could therefore be used to support or falsify the model.

REFERENCES

- Ariely, D., Gneezy, U., Loewenstein, G., and Mazar, N. (2009). Large stakes and big mistakes. *Rev. Econ. Stud.* 76, 451–469. doi: 10.1111/j.1467-937X.2009.00534.x
- Baldassarre, G., and Mirolli, M. (eds.). (2013). *Intrinsically Motivated Learning in Natural and Artificial Systems*. Berlin: Springer-Verlag. doi: 10.1007/978-3-642-32375-1
- Baranes, A., and Oudeyer, P.-Y. (2009). R-iac: robust intrinsically motivated exploration and active learning. *IEEE Trans. Auton. Mental Dev.* 1, 155–169. doi: 10.1109/TAMD.2009.2037513
- Bolado-Gomez, R., and Gurney, K. (2013). A biologically plausible embodied model of action discovery. *Front. Neurorobot.* 7:4. doi: 10.3389/fnbot.2013.00004
- Caluwaerts, K., Staffa, M., N'Guyen, S., Grand, C., Dolle, L., Favre-Felix, A., et al. (2012). A biologically inspired meta-control navigation system for the psikharpx rat robot. *Bioinspir. Biomim.* 7:025009. doi: 10.1088/1748-3182/7/2/025009
- Daw, N. D., Niv, Y., and Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat. Neurosci.* 8, 1704–1711. doi: 10.1038/nn1560
- Dearden, R., Friedman, N., and Russell, S. (1998). “Bayesian Q-learning,” in *Proceedings of the Fifteenth National Conference on Artificial Intelligence (AAAI)* (Menlo Park, CA: American Association for Artificial Intelligence), 761–768.
- Deci, E. L. (1971). Effects of externally mediated rewards on intrinsic motivation. *J. Pers. Soc. Psychol.* 18, 105–115. doi: 10.1037/h0030644
- Deci, E. L., Koestner, R., and Ryan, R. M. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychol. Bull.* 125, 627–668. doi: 10.1037/0033-2950.125.6.627
- Deci, E. L., and Ryan, R. M. (1985). *Intrinsic Motivation and Self-Determination in Human Behavior*. New York, NY: Plenum Press. doi: 10.1007/978-1-4612-2271-7
- Erez, M., Gopher, D., and Arzi, N. (1990). Effects of goal difficulty, self-set goals, and monetary rewards on dual task performance. *Organ. Behav. Hum. Decis. Process.* 47, 247–269. doi: 10.1016/0749-5978(90)90038-B
- Grzyb, B. J., Boedecker, J., Asada, M., del Pobil, A. P., and Smith, L. B. (2011). “Between frustration and elation: sense of control regulates the intrinsic motivation for motor learning,” in *AAAI Workshop on Lifelong learning* (San Francisco, CA).
- Holland, P. C. (2004). Relations between pavlovian-instrumental transfer and reinforcer devaluation. *J. Exp. Psychol. Anim.* Behav. Processes 30, 104–117. doi: 10.1037/0097-7403.30.2.104
- Kahneman, D., and Frederick, S. (2002). “Representativeness revisited: attribute substitution in intuitive judgment,” in *Heuristics and Biases: The Psychology of Intuitive Judgment*, eds T. Gilovich, D. Griffin, and D. Kahneman (Cambridge: Cambridge University Press), 49–81. doi: 10.1017/CBO9780511808098.004
- Keramati, M., Dezfouli, A., and Piray, P. (2011). Speed/accuracy trade-off between the habitual and the goal-directed processes. *PLoS Comput. Biol.* 7:e1002055. doi: 10.1371/journal.pcbi.1002055
- Killcross, S., and Coutureau, E. (2003). Coordination of actions and habits in the medial prefrontal cortex of rats. *Cereb. Cortex* 13, 400–408. doi: 10.1093/cercor/13.4.400
- Lepper, M., Greene, D., and Nisbett, R. (1973). Undermining children's intrinsic interest with extrinsic

4.3.1. Scope of motivation

In section 3.2.3 we found that in order to reproduce the empirical effects on activity, we have to assume that intrinsic motivation is local in scope rather than propagating all the way through the model. To our knowledge, no studies regarding the scope of intrinsic rewards exist, and it remains unclear how one could test such an aspect empirically. A possible approach could be to have individuals perform a creative task before introducing an unexpected event into the process. In two conditions to be compared, this event should either be immediately reproducible by the subject, say by pressing a previously unavailable button, or require a long sequence of actions to bring about. By observing whether the longer sequence causes less exploration in its direction or not, it should be possible to confirm or falsify our locality assumption.

4.3.2. Promised rewards

In section 3.3.2, we adopted the hypothesis of Niv et al. (2007) that expected rewards are explicitly encoded in the striatal system through tonic dopamine levels. In the framework of our model, assuming that the average reward \bar{R} encodes expectations leads to a system behavior that matches empirical findings by Ariely et al. (2009). Our model therefore supports the prediction that promises of rewards should indeed increase dopamine levels. However, the dopamine level theory is untested thus far, and would require an empirical study to confirm. Furthermore, a more detailed examination how promises of varying degrees influence behavior would be in order.

ACKNOWLEDGMENTS

The article processing charge was funded by the German Research Foundation (DFG) and the Albert Ludwigs University Freiburg in the funding program Open Access Publishing. We thank Nathaniel Daw for his support and for providing valuable references.

- rewards: a test of the “overjustification” hypothesis. *J. Pers. Soc. Psychol.* 28, 129–137. doi: 10.1037/h0035519
- Mannor, S., Simester, D., Sun, P., and Tsitsiklis, J. N. (2004). “Bias and variance in value function estimation,” in *Proceedings of the Twenty-First International Conference on Machine Learning, ICML* (New York, NY: ACM).
- McGraw, K. O., and McCullers, J. C. (1979). Evidence of a detrimental effect of extrinsic incentives on breaking a mental set. *J. Exp. Soc. Psychol.* 15, 285–294. doi: 10.1016/0022-1031(79)90039-8
- Mirolli, M., Santucci, V. G., and Baldassarre, G. (2013). Phasic dopamine as a prediction error of intrinsic and extrinsic reinforcements driving both action acquisition and reward maximization: a simulated robotic study. *Neural Netw.* 39, 40–51. doi: 10.1016/j.neunet.2012.12.012
- Niv, Y., Daw, N. D., Joel, D., and Dayan, P. (2007). Tonic dopamine: opportunity costs and the control of response vigor. *Psychopharmacology* 191, 507–520. doi: 10.1007/s00213-006-0502-4
- Oudeyer, P.-Y., and Kaplan, F. (2007). What is intrinsic motivation? A typology of computational approaches. *Front. Neurorobot.* 1:6. doi: 10.3389/neuro.12.006.2007
- Oudeyer, P.-Y., Kaplan, F., and Hafner, V. F. (2007). Intrinsic motivation systems for autonomous mental development. *IEEE Trans. Evol. Comput.* 11, 265–286. doi: 10.1109/TEVC.2006.890271
- Ryan, R. M., and Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *Am. Psychol.* 55, 68–78. doi: 10.1037/0003-066X.55.1.68
- Schembri, M., Mirolli, M., and Baldassarre, G. (2007). “Evolving internal reinforcers for an intrinsically motivated reinforcement-learning robot,” in *Proceedings of the 6th IEEE International Conference on Development and Learning*, eds Y. Demiris, D. Mareschal, B. Scassellati, and J. Weng (London: Imperial College), E1–E6. doi: 10.1109/DEVLRN.2007.4354052
- Schmidhuber, J. (1991a). “Curious model-building control systems,” in *Proceedings of the International Joint Conference on Neural Networks*. Vol. 2 (Singapore: IEEE), 1458–1463. doi: 10.1109/IJCNN.1991.170605
- Schmidhuber, J. (1991b). “A possibility for implementing curiosity and boredom in model-building neural controllers,” in *Proceedings of the International Conference on Simulation of Adaptive Behavior: From Animals to Animats*, eds J. A. Meyer and S. W. Wilson (Cambridge, MA: MIT Press/Bradford Books), 222–227.
- Schmidhuber, J. (2008). “Driven by compression progress,” in *Knowledge-Based Intelligent Information and Engineering Systems (KES-2008)*, volume 5177 of *Lecture Notes in Computer Science*, eds I. Lovrek, R. J. Howlett, and L. C. Jain (Berlin: Springer-Verlag), 11.
- Schmidhuber, J. (2010). Formal theory of creativity, fun, and intrinsic motivation (1990–2010). *IEEE Trans. Auton. Mental Dev.* 2, 230–247. doi: 10.1109/TAMD.2010.2056368
- Singh, S., Barto, A., and Chentanez, N. (2005). “Intrinsically motivated reinforcement learning,” in *Advances in Neural Information Processing Systems 17: Proceedings of the 2004 Conference*, eds L. K. Saul, Y. Weiss, and L. Bottou (Cambridge, MA: The MIT Press).
- Skinner, B. F. (1953). *Science and Human Behavior*. New York, NY: Macmillan.
- Sutton, R., and Barto, A. G. (1998). *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press.
- White, R. W. (1959). Motivation reconsidered. *Psychol. Rev.* 66, 297–333. doi: 10.1037/h0040934
- Yin, H. H., Ostlund, S. B., Knowlton, B. J., and Balleine, B. W. (2005). The role of the dorsomedial striatum in instrumental conditioning. *Eur. J. Neurosci.* 22, 513–523. doi: 10.1111/j.1460-9568.2005.04218.x

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 05 July 2013; accepted: 23 September 2013; published online: 16 October 2013.

Citation: Boedecker J, Lampe T and Riedmiller M (2013) Modeling effects of intrinsic and extrinsic rewards on the competition between striatal learning systems. *Front. Psychol.* 4:739. doi: 10.3389/fpsyg.2013.00739

This article was submitted to Cognitive Science, a section of the journal *Frontiers in Psychology*.

Copyright © 2013 Boedecker, Lampe and Riedmiller. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Keep focussing: striatal dopamine multiple functions resolved in a single mechanism tested in a simulated humanoid robot

Vincenzo G. Fiore^{1*}, Valerio Sperati², Francesco Mannella², Marco Mirolli², Kevin Gurney³, Karl Friston¹, Raymond J. Dolan¹ and Gianluca Baldassarre²

¹ Wellcome Trust Centre for Neuroimaging, Institute of Neurology, University College London, London, UK

² Laboratory of Computational Embodied Neuroscience, CNR, Istituto di Scienze e Tecnologie della Cognizione, Roma, Italy

³ Adaptive Behaviour Research Group, Department of Psychology, University of Sheffield, Sheffield, UK

Edited by:

Eddy J. Davelaar, Birkbeck College, UK

Reviewed by:

Dimitris Pinotsis, University College London, UK

Jennifer Lewis, University of Sheffield, UK

Dietmar Heine, University of Birmingham, UK

***Correspondence:**

Vincenzo G. Fiore, Wellcome Trust Centre for Neuroimaging, Institute of Neurology, University College London, 12 Queen Square, London WC1N 3BG, UK
e-mail: vincenzo.g.fiore@gmail.com

1. INTRODUCTION

Distinct functions are ascribed to striatal dopamine (DA) in relation to the type of outflow (tonic/phasic) expressed by this neuromodulator and the experimental context. “Tonic” DA release is caused by the removal of inhibitory constraints affecting spontaneously active DAergic neurons (Floresco et al., 2003; Grace et al., 2007). This low frequency mode of DA activation is considered as encoding average rewards (Niv et al., 2007; Beierholm et al., 2013), the presence of stressors (Cabib and Puglisi-Allegra, 2012) or novel stimuli (Lisman and Grace, 2005), and more recently as an indicator of precision of prior beliefs (Friston et al., 2012). As far as its function, tonic DA is mainly investigated for its effects on motor control: one influential account posits a role in mediating the vigor with which a subject pursues desired outcomes (Niv et al., 2007) which might be limited to approach strategies (Guitart-Masip et al., 2012). This overlaps with a proposed role in mediating the disposition to exert and sustain effort in pursuing a goal (Salamone et al., 2003; Salamone and Correa, 2012) and incentive salience in motivation or “wanting” (Berridge and Robinson, 1998; Peciña et al., 2003). Recent human evidence has also suggested a role attaining a balance between model free and model-based behaviors (Wunderlich et al., 2012), a formulation consistent with models of habitual versus goal control in Parkinson disease (Redgrave et al., 2010) and with DA’s established role in reasoning, cognitive flexibility, planning, and working memory (Montague et al., 2004; Cools and D’Esposito, 2011).

Phasic DA release results from a direct glutamatergic excitation of DAergic neurons (Floresco et al., 2003). There is substantial

The effects of striatal dopamine (DA) on behavior have been widely investigated over the past decades, with “phasic” burst firings considered as the key expression of a reward prediction error responsible for reinforcement learning. Less well studied is “tonic” DA, where putative functions include the idea that it is a regulator of vigor, incentive salience, disposition to exert an effort and a modulator of approach strategies. We present a model combining tonic and phasic DA to show how different outflows triggered by either intrinsically or extrinsically motivating stimuli dynamically affect the basal ganglia by impacting on a selection process this system performs on its cortical input. The model, which has been tested on the simulated humanoid robot iCub interacting with a mechatronic board, shows the putative functions ascribed to DA emerging from the combination of a standard computational mechanism coupled to a differential sensitivity to the presence of DA across the striatum.

Keywords: basal ganglia, dopamine, selection, novelty, iCub, intrinsic motivation

agreement these short burst firings play a key role in triggering learning processes, but the exact information they convey is disputed. The main proposal is that DA bursts report a reward prediction error resulting in reinforcement learning, a key element in behavior that leads to reward maximization (Sutton and Barto, 1998; Schultz, 2007). However, phasic DA is also considered as implicated in signaling saliency (Redgrave et al., 1999b) and agency-related novelty (Redgrave and Gurney, 2006; Redgrave et al., 2008).

Whether DA is considered as signaling the presence of unexpected or novel stimuli and independently of their association with the agent’s actions or priors, there exists a strong relation between DA and the broad category of intrinsically motivating stimuli. These are motivations guiding learning in the absence of primary “extrinsic” rewards such as food, water, and pain, and are directed to acquire knowledge and skills exploitable in later stages (Ryan and Deci, 2000; Baldassarre and Mirolli, 2013; Mirolli et al., 2013). The key feature of these motivations relies in the optimization of the information flow (Tishby and Polani, 2011), narrowing the amount of information that needs to be processed and motivating risky, but potentially fruitful, explorations in a changing environment (Kakade and Dayan, 2002; Ranganath and Rainer, 2003; Kaplan and Oudeyer, 2007; Düzel et al., 2010).

DAergic neurons are localized in a restricted brain region mainly the ventral tegmental area (VTA) and substantia nigra pars compacta (SNpc). By contrast, its targets, including the striatal region, are broad and heterogeneous. This is often seen as suggesting that DA cannot encode fine grain information and this lack of target specificity hints that its effects may be the

expression of a coarse influence (Schultz, 2007). Among DA principal projection targets is the striatum, a component in a complex circuitry involving the substantia nigra pars reticulata (SNr), the globus pallidus (GP) and the sub-thalamic nucleus (STN) that together form the basal ganglia. These nuclei are connected to the cortex via the thalamus to create parallel reentrant loops, where motor, associative, and ventral (limbic) cortices project to their specific target compartments in the striatum—respectively putamen (Put), caudate (Cau), and nucleus accumbens (NAcc)—(Alexander et al., 1986; Haber et al., 2000; Utter and Basso, 2008; Miyachi, 2009). With minor exceptions, these loops show qualitatively similar internal structure across functional areas (Nakano, 2000; Redgrave et al., 2010). The features characterizing this circuitry have led researchers to ascribe two functions to the basal ganglia: first, as responsible for action selection modulated by tonic DA outflow (Redgrave et al., 1999a), and, second, as mediator of reinforcement learning triggered by phasic DA via instrumental conditioning and novelty detection (Schultz, 2006). Thus, current theories highlight a neuromodulatory gain control and action selection role for DA or, alternatively, focus on its role in mediating the synaptic plasticity that underlies learning. Our approach rests upon coupling these two roles so that action selection and learning become an integral part of learning how to select actions. We will see later that this involves a closed causal chain involving the dopaminergic modulation of cortical plasticity and the cortical drive of phasic and tonic DAergic responses.

The core of our proposal is a new integrated hypothesis of the interaction between DA and cortical-striatal circuitry. In particular, we propose that DA's putative functions result from the combination of a differential sensitivity characterizing striatal subregions and the ability of DA to dynamically modulate a competition taking place within different basal ganglia nuclei. The present models show how DA affects the gain of a striato-cortical loop, altering the range of inputs capable of triggering a selection, the time required to perform a selection, and the ability of the system to persevere in a selection despite changes in the input. This mechanism is coupled with a differential sensitivity each part of the striatum exhibits to DA levels. This hypothesis is consistent with data describing the distribution of DA receptors in the striatum (Beckstead et al., 1988; Piggott et al., 1999) and it enables the agent to switch between behavioral strategies depending on the type of motivating stimuli perceived.

To support our hypothesis, we first simulate the activity of a single striato-cortical loop providing it an external arbitrary input and recording the way its processes are modified by the different outflows of DA. Secondly, we present a more complex model grounded on three striato-cortical loops, interconnected via the cortex, respectively for the control/selection of: arm actions (Put and pre-motor cortex, PMC), attention/associative processes for the selection of eye gaze (Cau and frontal eye field, FEF), and executive control for goal-directed behavior (NAcc and prefrontal cortex, PFC).

Both models are used in a series of simulated embodied tests performed on the humanoid robot iCub (Metta et al., 2010). The single loop model shows how increasing DA outflow enhances the probability of performing any selection (akin to action vigor)

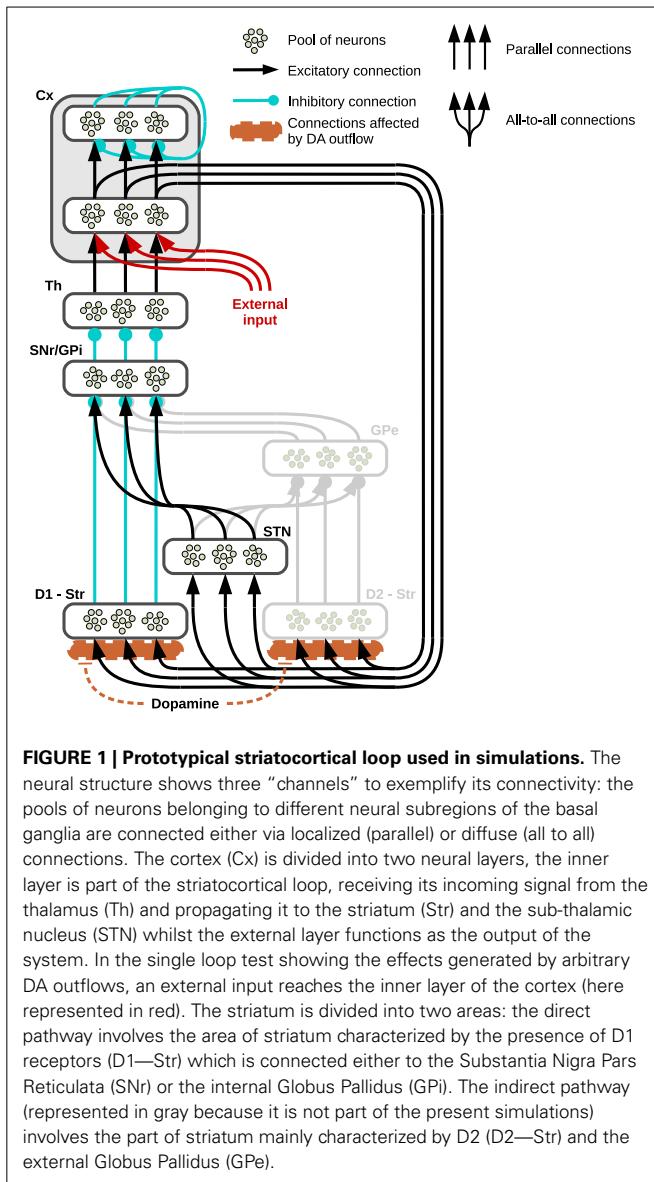
and leads to an increased perseverance of the selection in the face of distractors and variable information from environment. The three-looped model is used to solve a task requiring sensory-driven and novelty-driven exploration of a device having buttons and lights (the *mechatronic board*, cf. Taffoni et al., 2013): the agent is required to learn via intrinsic motivation (i.e., unexpected visual stimuli, Reed et al., 1996) and to exploit the acquired associations when extrinsic rewards appear in the environment. The next section will describe the details of the parts of the basal ganglia we have focussed on, neglecting others to simplify the overall complexity of the biological system the models refers to. Despite these simplifications, we think the results of the tests show that the DA-based mechanisms illustrated above can play several important adaptive functions such as the guidance of sensory- and novelty-driven exploration, the exploitation of goal-directed (model-based) action-outcome associations, and the saving of energy (rest) when no motivating stimuli are perceived.

2. MATERIALS AND METHODS

2.1. BASAL GANGLIA: ANATOMY AND CIRCUITRY

The multifunctional role ascribed to the action of DA within the striatum renders it unsurprising that the basal ganglia are themselves implicated in guiding perception, attention, learning, and memory processes, beside motor control. Both empirical evidence (Mink, 1996; Redgrave et al., 1999a; Grillner et al., 2005; Hikosaka, 2007) and computational modeling (see Humphries et al., 2006; Prescott et al., 2006; Baldassarre et al., 2012; Humphries et al., 2012, for the most closely related to the present model and Frank, 2011 for a general review) converge on the idea that a core element of basal ganglia function involves removal of tonic inhibition so as to realize a selection of its input.

The basal ganglia receive massive input from most regions of cortex and provide a processed output to the thalamus, which closes the loop via reconnection back to the cortex. The circuitry characterizing the cortico-thalamic connection is also rather complex: the thalamus reaches layer IV of the cortex and this reaches the striatum via layers III and V whilst another loop involving directly thalamus and cortex is closed via layer VI (Douglas and Martin, 2004; da Costa and Martin, 2010). For the purpose of this study, the architecture will capture only the features characterizing specific parts of the basal ganglia relevant to the objectives of this work, leaving aside the complex interaction involving the other two main actors in this loop, namely cortex and thalamus (see section 4 for further details). One of these essential features is illustrated in **Figure 1**, which shows the parallel “channels” of neural populations characterizing a striatocortical loop (Alexander et al., 1986; Alexander and Crutcher, 1990; Gurney et al., 2001a,b): the striatum receives its localized input directly from the cortex and it propagates this signal via two distinct pathways, each originating in a subregion characterized by the presence of specific DA receptors. The first of these two striatal subregions shows a higher concentration of D1 receptors (having excitatory effect) and directly connects to the SNr (when considering the NAcc) and the internal part of the GP (Gpi, when considering the Cau and Put), forming the so-called direct pathway; the second subregion is characterized by greater concentration of D2 receptors (having an inhibitory effect) and its signals



reach the SNr/GPi via a double inhibition involving the external Globus Pallidum (GPe), the so-called indirect pathway. Finally, a cortical input also reaches the STN which is connected directly to the SNr and GP via diffuse excitatory connections referred to as the hyperdirect pathway.

Parallel inhibitory channels of neural populations run through the whole loop, in both the direct and indirect pathways, as opposed to the diffuse excitatory connections between STN and SNr/GP. This structure results in a functional double competition between two regions preserving segregated activations and the region providing a diffuse undifferentiated signal: the former regions convey information about the values of each separate component of the input, whereas the latter conveys non-specific information about the general intensity of the incoming stimuli as a whole (Frank, 2006; Frank et al., 2007).

Assuming the input provided by the cortex already encodes the value or salience of the stimuli (Samejima et al., 2005; Lau

and Glimcher, 2008; Kimchi and Laubach, 2009; FitzGerald et al., 2012; Znamenskiy and Zador, 2013), the input nuclei of the three pathways receive and process these saliences in a continuous self-feeding process mediated by the presence of a closed loop: depending on the relative strength of activity in these pathways, the basal ganglia eventually alter these values preserving, increasing or suppressing the differences encoded. This process is mediated by the tonic inhibitory activity of the SNr/GPi—the output nucleus of the basal ganglia—whose channels can be selectively inhibited so as to release the corresponding population of neurons in the thalamus and resulting in a gating effect (Chevalier and Deniau, 1990; Gurney et al., 2001a,b). Most of this tonic activity is provided by the hyperdirect pathway which therefore concurs in reducing the chances that any of the channels in the SNr might be inhibited; on the contrary, the direct and indirect pathways compete in establishing which of the SNr/GPi channel has to be inhibited, the former favoring the strongest cortical inputs whereas the latter favors the weakest.

In the present study we are mainly interested in testing the effects on behavior due to an increase in DA outflows. Thus, we have simplified the structure of basal ganglia by relying on a model that focusses on the competition implemented by direct and hyperdirect pathways alone (**Figure 1** shows the regions whose activity has not been simulated in light gray). This simplification is justified assuming that, due to the presence of the D2 receptors, increasing DA release causes the indirect pathway to decrease its activity, therefore—in a computational perspective—it diminishes its effect on the whole system, allowing the D1-related direct pathway to have a major role in the selections (Humphries et al., 2012). This choice is also consistent with data and models identifying indirect pathway structures as responsible for “No-Go” that is negatively correlated with increases of DA release due to high concentrations of D2 receptors (Frank et al., 2004; Surmeier et al., 2007; Frank, 2011; Guitart-Masip et al., 2012): the present models rely on a simplified structure which can be considered nonetheless accurate in analyzing most of the behaviors connected with high DA outflows and “Go” choices.

2.2. THE COMPUTATIONAL MODEL

The neural systems used for both simulation and embodied tests were developed with C++ libraries: these were tested for the first time in Baldassarre et al. (2012) and have been modified to deal with the new requirements concerning the neural architecture and the mechanics involving the simulated DA. The basic building block of the models is a *leaky integrator unit* defined by a continuous-time differential equation that simulates mean activity of a whole neural area or pool of neurons. This is a standard tool in firing rate models (Dayan and Abbott, 2005), modified to include the effects of the DA neuromodulation as follows:

$$\tau_g \dot{u}_j = -u_j + b_j + (\epsilon + \lambda d) \Sigma_i w_{ji} y_i \quad (1)$$

where τ_g is a time constant (related to the nucleus or group, g , of units to which j belongs), u_j is the activation potential of unit j , b_j is the basal activation of such unit (if any), w_{ji} represents the connection weight between input unit i and unit j , and y_i the activation of input unit i .

To include the DAergic modulation, we assume DA enhancement of the signal reaching a target area can be simulated via a multiplicative effect: this is a standard computational strategy in simulating D1 specific effects (Fellous and Linster, 1998; Durstewitz, 2009) and is realized through the parameter d , representing the amount of DA released, and the coefficients ϵ and λ , respectively for the strength of the input independent of the presence of DA and the multiplicative effect DA exerts on the same input. These two coefficients have been set to $\epsilon = 1$ and $\lambda = 0$ for all the units which are not affected by the DA release in the simulations, and $0 < \epsilon < 1$ and $\lambda > 0$ for the remaining units: besides the striatum, the three-looped model (see **Figure 2**) also shows the hippocampal simulated layer as being affected by the presence of DA in the way described here.

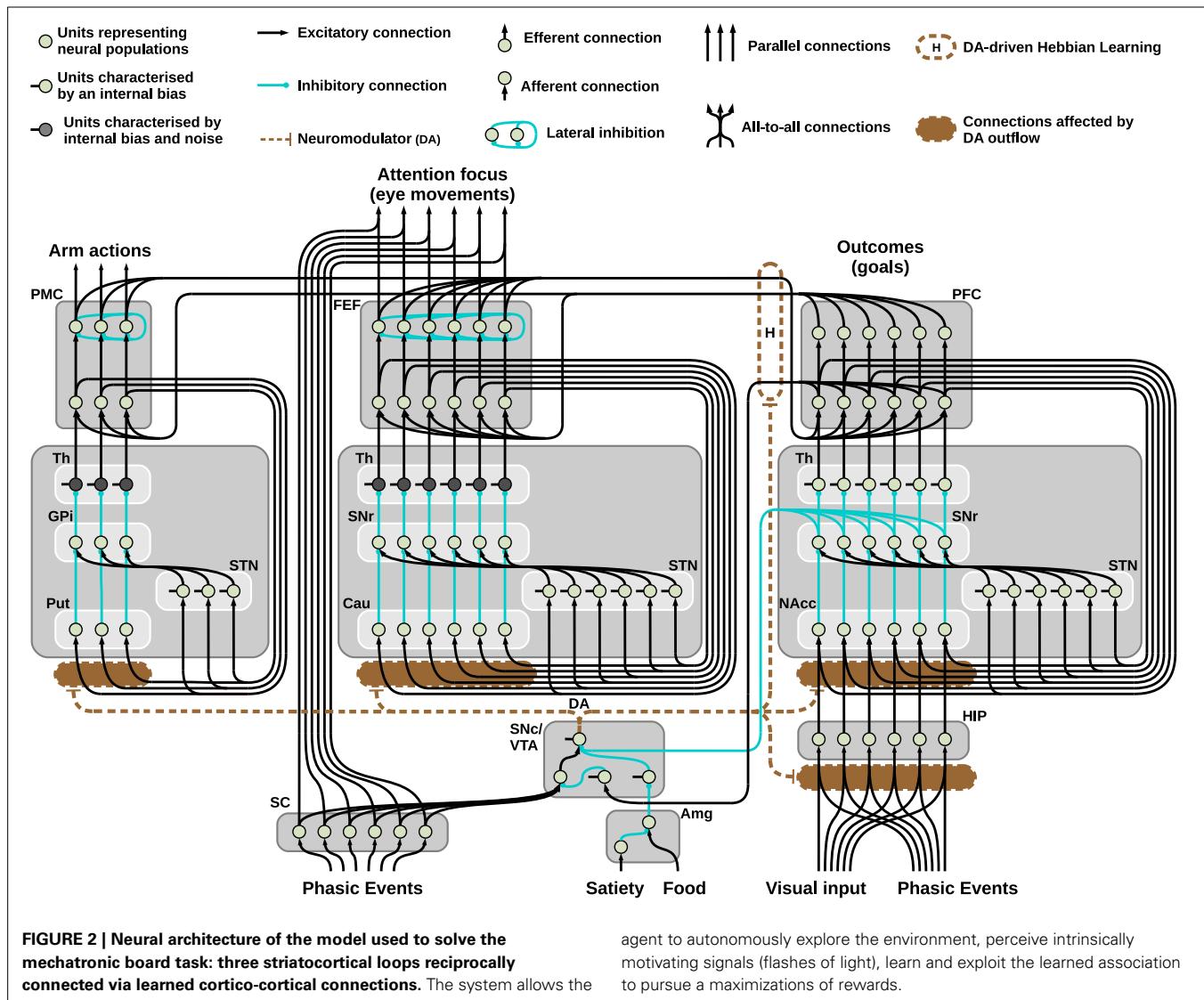
Equation (1) describes the activation potential of the units in the neural models where Equation (2) is a positive saturation transfer function defining the final activation of these units: the activity of all units here described is simulated relying on these

two equations. The transfer function is defined as follows:

$$y_j = [\tanh(\alpha_g(u_j - \theta_g))]^+ \quad (2)$$

where $\tanh(\cdot)$ is the hyperbolic tangent function, α_g is a constant defining the slope of the hyperbolic function (per group), θ_g is a threshold parameter (per group) and $[x]^+ = 0$ if $x \leq 0$ and $[x]^+ = x$ if $x > 0$. Notice that $0 < y_j < 1$ for all j . Aside from the layers simulating the activity of the cortex, the threshold is always set to $\theta = 0$: the cortical transfer function has been thresholded so that units activations are zero unless their corresponding activation potential exceeds its layer specific threshold (in the single loop simulation, these are set to 0.6 for the inner cortical units and 0.8 for external cortical units).

Finally, the three-looped model shown in **Figure 2** consists of three separated loops for manipulation, attention, and executive control: respectively the dorsolateral, dorsomedial, and ventral striatocortical loop. During the task, these systems establish



cortico-cortical connections that reciprocally bias the selection performed thanks to a Hebbian learning process guided by the presence of phasic DA. The equation describing this learning is as follows:

$$\Delta w_{ji} = \eta_{ctx} g_j y_i (\hat{w}_{ctx} - w_{ji}) [d - \zeta]^+ \quad (3)$$

where y_i and g_j represent the activities of the connected units (belonging respectively to the cortical external layer in the pre-synaptic loop and the cortical inner layer in the post-synaptic loop), w_{ji} is the connection weight between y_i and g_j , η_{ctx} is a learning rate and \hat{w}_{ctx} is a maximum value reachable by w_{ji} . The neuromodulator is here thresholded: $[d - \zeta]^+$ represents the amount of DA required to overcome a threshold ζ , where $[.]^+$ is defined as in Equation (2). This threshold is set higher than any tonic outflow variation, therefore allowing learning processes only in presence of high peaks of DA corresponding to phasic activations.

The DA bursts required for LTP Hebbian learning are triggered by sudden luminance variances perceived by the system/agent via the superior colliculus (SC): this region provides fast and strong signals to the DAergic units, which result in the simulated DA bursts (resembling the actual connectivity and function as described by Redgrave and Gurney, 2006). Both tonic and phasic DA releases are simulated with one component representing the overall activity of both the VTA and the SNc: the activity of the single DAergic unit is controlled by both excitatory and (tonically active) inhibitory units.

In order to simulate the presence of an agency-related predictor, the same learning process expressed in Equation (3) is also used to establish excitatory connections between the inner cortical layer of the simulated PFC, part of the ventral loop, and an interneuron unit in the DAergic area: this direct connection simplifies the actual pathway responsible for this signal control functioning, which may involve the lateral habenula (Hikosaka et al., 2008). This learning process, triggered by the presence of DA bursts, eventually leads to suppression of phasic DA responses: the agent relies on the acquired cortico-cortical associations between specific combination of attentional/motor selections and PFC activity triggered by the perception of motivating stimuli to provide the ventral loop with the required information about the proximal cause of any experienced motivating stimuli. Since motor and attentional selections temporally precede the stimulus, once the association is learnt, this information is sufficient to cause activation of an inhibitory unit in the DAergic area (via PFC) before the actual stimulus takes place. As a result, an action causing unexpected changes in the environment, such as luminance variance in the present task, will trigger DA bursts that engage learning processes among cortical regions and between the PFC and the DAergic area. However, if manages to successfully repeat the correct action on the proper target, the resulting change in the environment will eventually become predicted, therefore preventing an input coming from the SC from triggering any more DA bursts.

The learned cortico-cortical connections among different striato-cortical loops are instances of inverse and forward models (Gurney et al., 2013). Inverse models implement here the links

between goal representations and action representations, important for the recall of actions on the basis of the pursued goals in goal-directed behavior. The forward models, instead, allow the anticipation of the accomplishment of a certain outcome when a certain action is performed.

This role of DA in the self-assembly or bootstrapping of intrinsically valuable sensorimotor sequences is reminiscent of early simulations of value-dependent learning using neuronally plausible models (Friston et al., 1994). In brief, the dopaminergic reinforcement of stimulus-response and response-stimulus links by DA depends upon phasic dopaminergic discharges. By introducing dopaminergic plasticity into the cortical projections eliciting these discharges, one introduces a circular causality, in which innately or intrinsically rewarding stimuli transfer their value to their sensory or motor precedents. This form of learning has formal links with actor-critic models in reinforcement learning, accounts for the transfer of phasic dopaminergic responses from unconditioned to conditioned stimuli and provides a physiologically grounded account of how sequences of exploratory or exploitative behavior emerge.

Among the remaining components of the model pictured in Figure 2, the hippocampus (HIP) is composed by a single layer of units encoding spatial representations: the activity of these units slowly decreases as a response to the incoming input. The slow decrease of the input (which starts from the maximum value of 1 to reach its minimum value of 0.1 in roughly 2 min) is determined by the time of exposure to the visual stimulus: this process simulates habituation to novel stimuli, leading to high responses of HIP to novel stimuli located in space (as it happens during visual exploration of a new environment) and low responses in presence of familiar items.

The projections of the HIP via the NAcc and the SNr to DAergic areas drives changes toward a tonic response mode of the simulated DAergic unit, which itself affects the activity of the HIP thus creating a loop. This circuitry is consistent with HIP connectivity and functioning (Grace et al., 2007) as the literature describes it as one of the major systems responsible for novelty detection and the related regulation of tonic DA release (Lisman and Grace, 2005; Düzel et al., 2010). The HIP is not the only part of brain that responds to novelty and habituates (see Ranganath and Rainer, 2003, for a review). However, coherently with the choice illustrated above on the HIP as the only source of novelty detection in the model, we included in the model only HIP habituation. This assumption was sufficient to have a brain mechanism performing novelty detection and habituation, and the consequent novelty based tonic DA regulation.

The simulated DAergic area is thus controlled by activity of SC (causing phasic DA bursts), the PFC (inhibiting the signal coming from the SC and suppressing DA bursts), SNr (responsible for tonic inhibitory control mainly due to the HIP) and finally a simplified amygdala (Amg): this component affects the activity of a tonically active interneuron in the DAergic area, resulting in strong increase in the DA outflow when a reward is perceived (i.e., simulating the perceived presence of food in one of the boxes).

Regarding the BG-cortical structures, the manipulation striato-cortical loop is characterized by three channels and

Table 1 | Table of essential parameters marking the difference among the three loops and the learning processes: the complete set of parameters is available for download (see instructions in the Supplementary Material).

Parameters	Attention	Arm action	Goal
λ	2.5	1.5	1
θ_g layer 1	0.4	0.6	0.1
θ_g layer 2	0.8	0.8	0.6
Lateral inhibitions	2	0.2	0
Noise in Th	20	30	0
Noise decay	1000	2000	0
Learning processes	(ζ coefficient)		
Cortico-cortical	0.2		
Predictor	0.00008		

The tuning has been carried out by comparison with behavioral results.

controls the arm in the robotic set-up allowing selection among three possible actions (one per channel). Both the attentional loop and the ventral loop have six channels: the first controls saccade among six possible locations in space and the second controls the selection of the desired outcome to pursue. The three loops are similar, showing differences in only a few key parameters: among these, it is important to stress the presence of random noise in the thalamic parts of the manipulation and attentional loops and the presence of a different value for the coefficient λ for each of the striatal layers. The noise, which is essential to perform random exploration, is smoothed using a leaky integrator (Equation 1) and therefore is controlled by two parameters, one for the strength of the input and one for the decay speed (see Table 1). The coefficient λ (Equation 1), on the other hand, simulates the differential sensitivity to the presence of DA characterizing different striatal regions. This differential sensitivity will be shown to be essential for endowing the system with a flexible behavioral expression and for avoiding multiple fixated selections.

The biological plausibility of this hypothesis is grounded on the known distribution of D1 receptors within the striatum: there is a gradient of D1 receptor density within each subregion, with the Cau and the NAcc having, respectively, the highest and the lowest concentrations (Beckstead et al., 1988; Piggott et al., 1999). Assuming a higher concentration of D1 makes a neural region more sensitive to any variation of DA outflow is consistent with the model computational requirements to solve the mechatronic board task. In the model, DA alters the gain in each of the three feedback loops, having in the attentional loop (involving the Cau) the most sensitive system, in the manipulation loop (involving the Put) mid sensitivity and in the executive control loop (involving the NAcc) the system that requires the most DA release to be activated.

2.3. ROBOTIC SETUP AND MECHATRONIC BOARD TASK

The *iCub*¹ is a humanoid robot whose dimensions resemble those of a 5 years old child. This robotic platform is characterized by

an high number of degrees of freedom (16 for each arm, 5 for the head-eyes, 3 for the torso), so it is particularly fit to deal with tasks involving “human-like” movements. The official simulator of the *iCub* has been used to run the experiments concerning vigor and the solution to the *mechatronic board task* (see Figure 3). A mechatronic board, described in Taffoni et al. (2013), has been simulated and employed as the test environment. In order to match the requirements of the three-looped neural system here described, three actions (namely “grab,” “wipe,” and “press”), have been implemented to move the robot left hand in different ways and positions. Any selected action is always performed on the target the *iCub* is looking at. Through its movements the robot can interact with the mechatronic board, triggering light-flashes (lasting 1 s) when the proper action is performed on one of the correct targets: the time required to complete an action varies between 2 and 3 s circa (0.5 for a saccade), depending on the starting position of the arm and the final target. Note that, despite the name, the actions “grab” and “wipe” denote simply dummy actions, i.e., actions with no consequence on the board.

The control works in continuous time reflecting the activity of the neural system, so that both the actions and the targets can be changed or stopped at any time. This feature allows the experimenter to add and relocate a reward in any of the accessible locations at any time whilst the robot is interacting with the environment. A link to a short movie showing the robot interacting with the actual mechatronic board is provided in the Supplementary Material.

The task the agent is dealing with is rather simple: it requires exploration of an unknown environment, learning of agency-related associations due to the presence of intrinsically motivating stimuli (light flashes) and recall/exploitation to pursue the maximization of extrinsically motivating rewards. The mechatronic board consists of three buttons and three transparent boxes (see Figure 3): when the correct action is performed on any button (press), the box opens and the associated light flashes. The agent is provided with a sufficient amount of time to freely explore its accessible environment. In a second phase of the task the environment is modified adding a visible reward (e.g., food) inside one box: to access the reward, the agent is required to recall the learned association and to perform the correct action causing the opening of the box. The task, which resembles the response pre-conditioning driven by neutral stimuli described by Reed et al. (1996), has already been solved using an early version of the three-looped model (Baldassarre et al., 2012): a comparison between the two versions of the model is provided in section 4.

3. RESULTS

3.1. INPUT DISCRIMINATION: EFFECT OF DA IN A SINGLE LOOP

To show the effects different outflows of DA have on the processes performed by the Basal Ganglia, several tests have been carried out on a three-channel loop as in Figure 1: the mean activity of pools of neurons has been simulated as in Equations (1, 2) and an arbitrary input lasting 6 min, consisting of a three-dimensional vector, has been set to reach the inner cortical layer of the cortex.

Figure 4 (left) shows values and variations of the input vector assigning a different color (blue, green, red) to each of the three-dimensions: the input changes five times during each test,

¹<http://www.icub.org/>

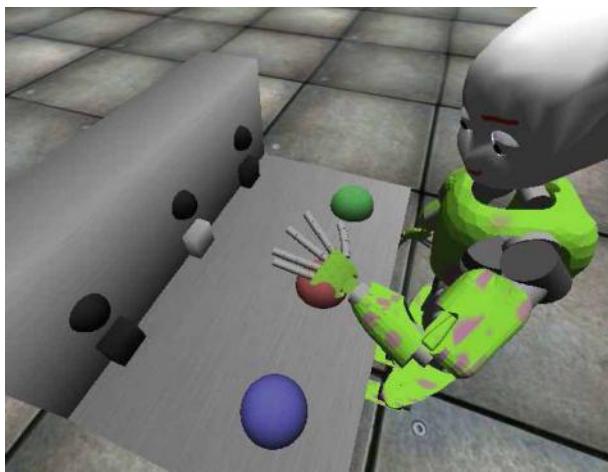


FIGURE 3 | The iCub in its simulated environment interacting with the mechatronic board. The image is captured whilst the robot is pressing the red button, triggering the corresponding light.

with a fixed interval of 1 min. The same color code has been used to represent activation of the corresponding channels as recorded in the inner layer of the cortex and depicted in the center column of **Figure 4**: any of the units in the cortical inner layer may independently overcome a threshold (0.8, pictured as a dotted line) characterizing the transfer function of the external layer of the cortex. The activity in the external layer of the cortex is stabilized by the presence of lateral inhibitions preventing this layer from exhibiting multiple channel activations (**Figure 4**, B/W colormaps).

Given the specific set of parameters characterizing this loop, the input has been chosen so that it shows two features: first it is insufficient by itself to cause activation of any unit in the cortical external layer in the baseline DA condition. Secondly, the changes in the components in the input vector alter both sparseness (mean interval) and the overall mean value. The tests show that, depending on the amount of DA reaching the striatum, it is useful to distinguish three conditions:

- (1) *Weak or scarce selection.* In this condition, the striatal-cortical loop requires an input which must be both strong and sparse in order to overcome the given threshold. Therefore, few selections are performed, and—because of strong correlation with input features—they are characterized by high instability, being abandoned as soon as either the intensity of the strongest stimulus decreases or any other stimulus increases its intensity. The first row in **Figure 4** (DA +15%) exemplifies this condition, showing only one activation over the threshold (fifth interval), despite the presence of stronger or equally valued stimuli in several other time intervals (i.e., third, fourth, and sixth).
- (2) *Enhanced discrimination.* The DA unbalances the competition between diffuse (STN) and localized (Str) signal processing, favoring the latter: this condition enables the loop to amplify the differences between stimuli with similar intensity

via accumulation of the strongest signal and suppression of the weaker ones. The time required to perform this process is directly correlated with the amount of DA (within a certain range, the higher the release, the faster the amplification, and thus the selection). Despite the fact the loop is still unable to discriminate between strong, closely related stimuli (e.g., sixth time interval), this condition is shown to be the most flexible to any change in the environment allowing, in most cases, quick switches in selection depending on the values encoded in the input. In particular, the comparison between the second and the third row (DA +18% +20%) illustrates the effect of accumulation granted by the loop and its timing: a higher level of DA allows the system to reach a homeostasis characterized by values which overcome the given threshold (time intervals 1, 2, 3, and 5 result in activations in the external layer of the cortex). Each time the input is propagated back from cortex to the striatum, the higher value encoded in the input grows: comparing the first two time intervals in these rows we notice that a slight increase in the DA outflow makes the input in the first interval cause an activation roughly 30 s in advance.

- (3) *Maintenance and disrupted selection.* The competition between localized and diffuse activation is strongly unbalanced in favor of the former: this allows the possibility of discrimination between closely related strong inputs (e.g., sixth time interval, condition DA +40%), but at the same time it makes any selection performed persistent so causing interference and delayed switch (first to second time interval, DA +25%), maintenance (first to second time interval, DA +30%) and eventually (if the DA further increases) multiple channel activation (third to sixth intervals, DA +50%). The system is now unable to respond quickly to changes in the stimuli unless they are characterized by strong values: any selection is preserved until either the DA outflow decreases or the input changes dramatically. A further increment of the DA outflow makes the maintenance effect so strong that multiple activations in the loop become more and more likely, disrupting a selection mechanism (which in the present model is preserved in the external layer only due to the action of the lateral inhibitions). Such a condition implies difficulties in adaptation to changes in the environment, but it can also be considered as the cause of a useful “focus effect” which allows the preservation of rewarding selections in the presence of noise or distractors. Indeed, the condition of maintenance may be reached both due to elevated tonic DA release and due to high frequency burst firings causing DA accumulation (Floresco et al., 2003): this phenomenon would therefore favor both the expression of incentive salience and learning processes granting the repetition of those selections that have proximally caused the increase in the DA outflow.

3.2. SIMULATION OF DA-DEPENDENT VIGOR

A second test has been carried out involving two segregated striatal-cortical loops, one characterized by six channels controlling attention via oculomotor selection (assuming the simplified environment of the mechatronic board showing only six cues to focus the attention on) and a second three channeled loop simulating

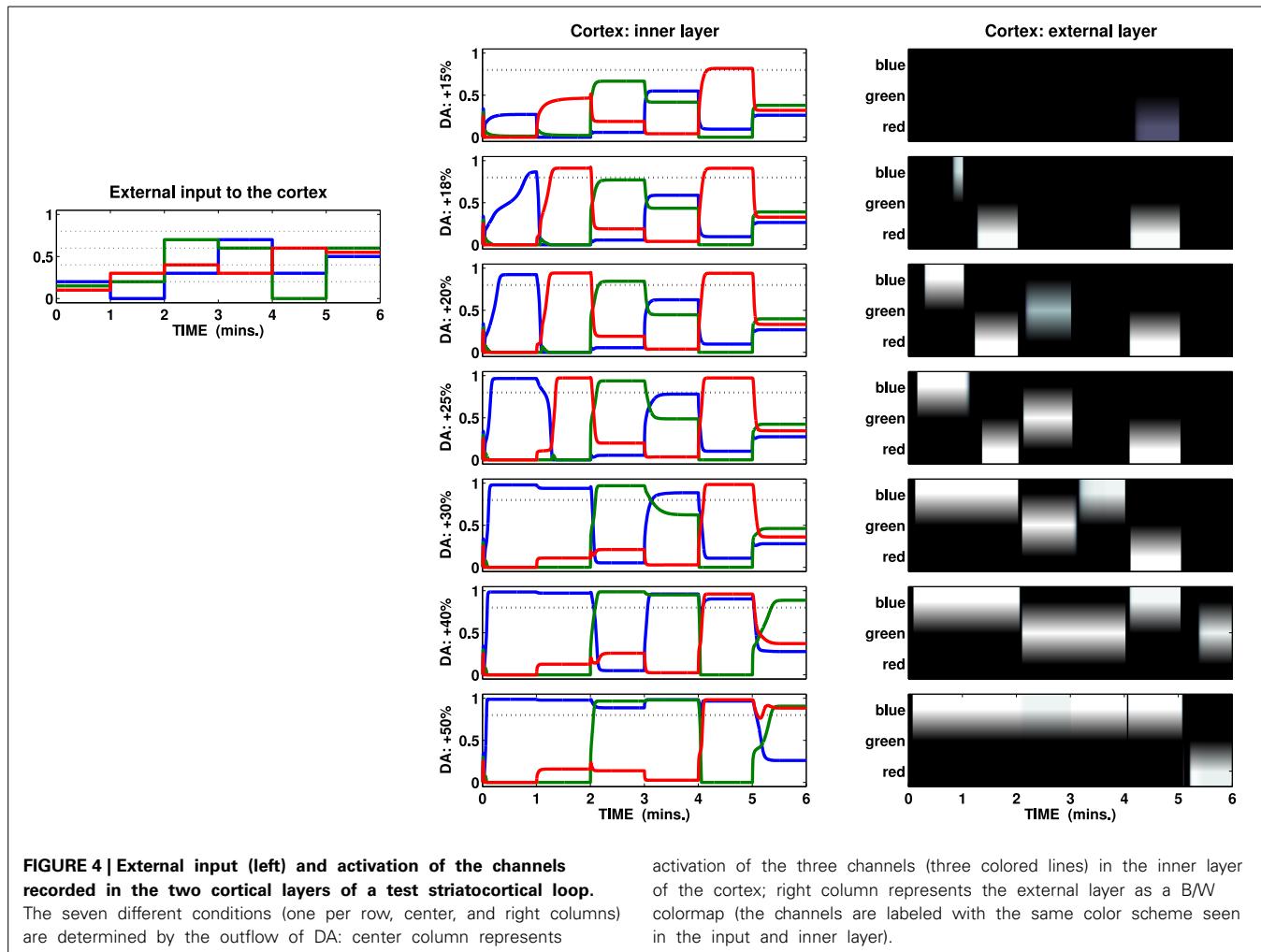


FIGURE 4 | External input (left) and activation of the channels recorded in the two cortical layers of a test striatocortical loop.

The seven different conditions (one per row, center, and right columns) are determined by the outflow of DA: center column represents

activation of the three channels (three colored lines) in the inner layer of the cortex; right column represents the external layer as a B/W colormap (the channels are labeled with the same color scheme seen in the input and inner layer).

the selection between three arm actions: in both cases, the loops do not receive any external input but they are provided with randomly generated noise in the thalamus. Changing every step, the noise results in a “random walk” eventually triggering random selection in the cortex. The choice of the thalamus as the locus for the random walk is justified by the reasoning that this area receives information from several cortical sources: this input is abstracted with the noise used in the model. In this respect, this noise should not be interpreted as local neural noise, but rather as the neural activity reaching the thalamus from different cortical areas and capable to overcome SNr/GPi inhibition, inducing exploration (Baldassarre et al., 2012). It is necessary to focus on a target and to select one of the arm-controlling channels to start executing any motor action: the agent requires a variable time of around 2–3 s to complete any hypothetical action on any selected target, so that both attentional and manipulation selections must be maintained for a sufficient amount of time. If the agent perseverates in its selections, the action is executed again on the same target, resulting in repetitions.

We ran several tests lasting 6 min on the iCub simulator changing the DA outflow (baseline, +20% and +40%) and recording the number of completed actions performed on any possible

target. The results are represented in Figure 5, which shows mean and standard error of completed actions—recorded in ten sample tests—in the three DA conditions and a sample test showing activations of external cortical layers in both the loops (B/W colormaps) and the corresponding actions performed in the three DA conditions.

These results are consistent with a known correlation between DA outflow and reward-related vigor (Niv et al., 2007; Beierholm et al., 2013) or incentive salience (Berridge and Robinson, 1998; Peciña et al., 2003), but they provide a new explanation of these behaviors. The number of completed actions increases significantly from an average of 0.6 (baseline) to 28.2 (DA +20%), reaching 91.8 (DA +40%) per test. This result is neither caused by any learning process nor it relies solely on the strengthening of the input due to DA multiplicative effect (e.g., Gurney et al., 2001b; Humphries et al., 2006): DA alters the gain of the loop thereby unbalancing the competition between the striatum and the STN, causing quick accumulation and selection at first (as seen in Figure 4, DA +20%) and then maintenance for a longer time (allowing repetitions, DA +40%).

These tests show that a widely known function ascribed to striatal DA can be produced by relying on a dynamic mechanism

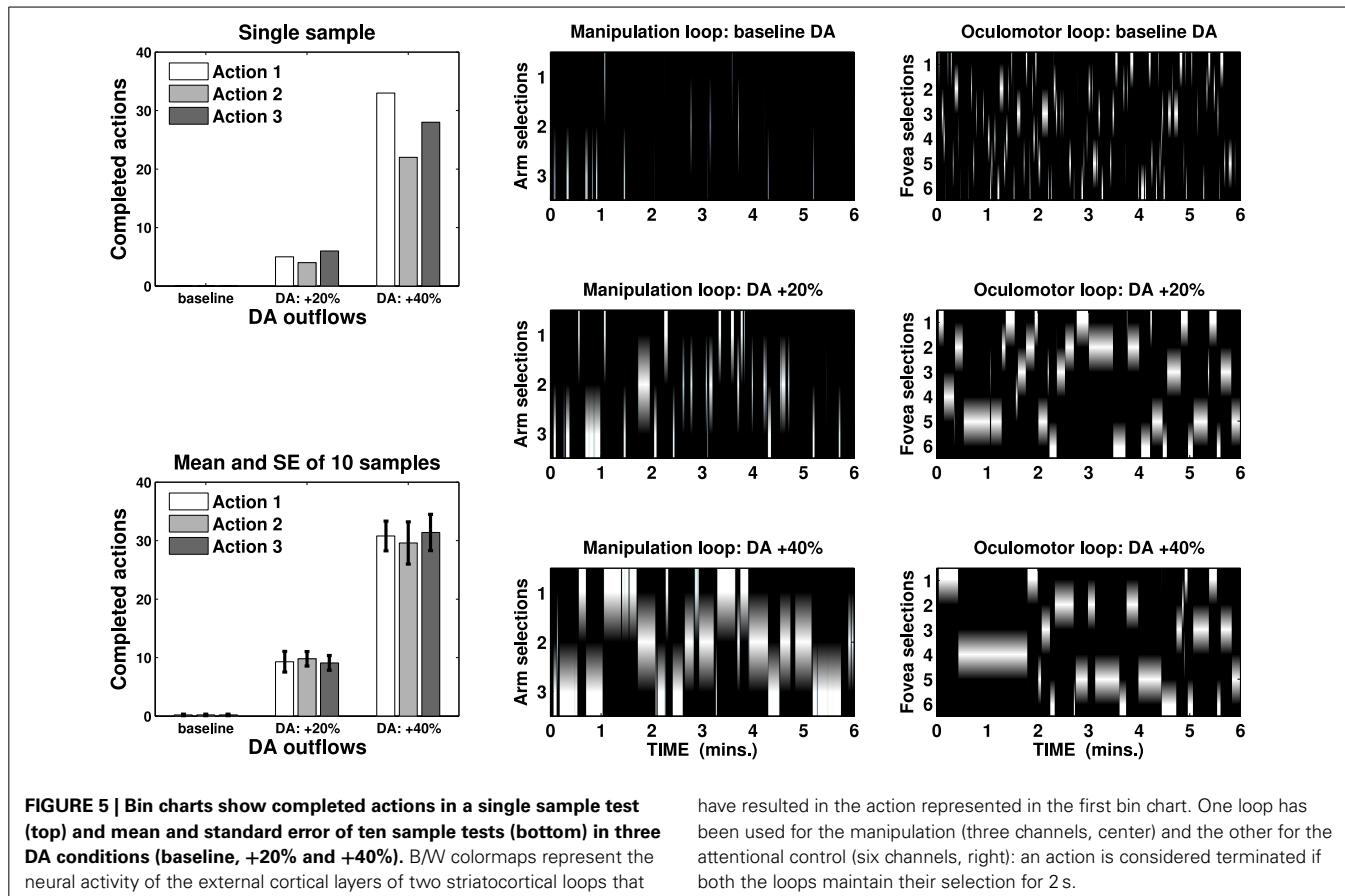


FIGURE 5 | Bin charts show completed actions in a single sample test (top) and mean and standard error of ten sample tests (bottom) in three DA conditions (baseline, +20% and +40%). B/W colormaps represent the neural activity of the external cortical layers of two striatocortical loops that

have resulted in the action represented in the first bin chart. One loop has been used for the manipulation (three channels, center) and the other for the attentional control (six channels, right): an action is considered terminated if both the loops maintain their selection for 2 s.

which allows the system to focus on a single rewarded selection as long as it is the cause of an increase of DA: the mechatronic board task exemplifies how this phenomenon both coexists and assists the standard computational role ascribed to DA as the trigger for learning processes.

3.3. THE SOLUTION TO THE MECHATRONIC BOARD TASK

To solve the task the three-looped model (see **Figure 2**) relies on the hypothesis that, due to a differential sensitivity in the striatum, the same DA outflow causes the manipulation loop to express the first behavior (weak or scarce selection) whereas the attentional loop expresses the second (enhanced discrimination). This is consistent with data in MPTP-induced Parkinsonian subjects associating low DA outflow with the absence of motor activity but slow oculomotor foveations (Hotson et al., 1986; Schultz et al., 1989; Hikosaka et al., 2000).

This differentiation makes the agent start a visual exploration of the environment whilst performing very few arm actions (as seen in **Figure 5**, B/W colormaps of baseline DA condition): as soon as a novel experienced cue is perceived then activity in the HIP triggers (via NAcc and SNr) an increase in the tonic release of DA, allowing the manipulation system to enter the condition of enhanced discrimination and forcing the attentional system to a state of maintenance (as seen in **Figure 5**, B/W colormaps of DA +20% condition). As a consequence, the agent starts executing on a single target several randomly selected actions: the

process stops when the HIP habituates to the perceived cue, restoring the usual outflow of DA, allowing the visual exploration to start again and reducing the number of action performed.

During this visual and motor exploration, the agent eventually focusses on any of the button cues: if the action “reach/press” is randomly selected and completed whilst on this target (the agent requires the usual 2–3 s of maintenance), the box associated with the pressed button opens and the corresponding light flashes. Sudden luminance changes are then perceived by the SC causing DA bursts which have a twofold effect: first, phasic DA is itself causative in a further tonic release of DA via HIP (which is highly sensitive to DA presence) therefore forcing maintenance in both manipulation and attentional loops and causing enhanced selection in the ventral loop (e.g., **Figure 7**, left column, 155–185 s interval). Secondly, phasic DA allows learning processes to take place, strengthening connection weights between different cortical layers and between the PFC and the DAergic area. This latter type of learning is responsible for the emergence of the agency-related predictor which results in an inhibition of DA bursts when specific motor/attention combinations are selected. As a consequence, attention is preserved on the target (the button) and the action (reach/press) is repeated until the DA bursts disappear because of the predictor, allowing the exploratory routine to restart.

The cortico-cortical learning, on the other hand, allows associating the selection of the reach/press action in the PMC and the

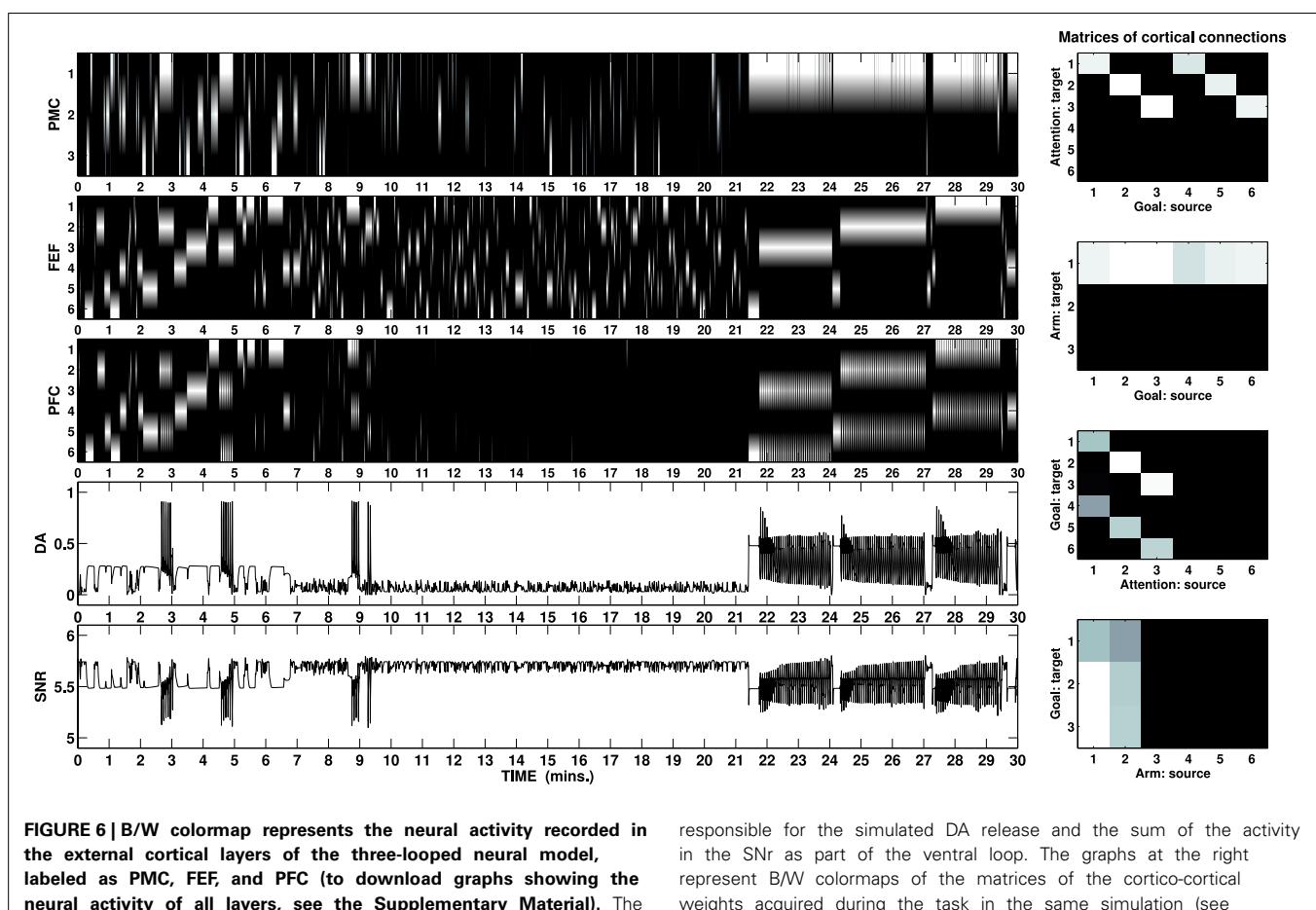
selection of attention on the button in the FEF with the channels activated at the same time in the PFC due to the activity of the HIP. The connections established between external PMC/FEF and internal layer of PFC are essential for the predictor, whereas the connections established between the external PFC and the internal layers of PMC and FEF are essential for the agent to express goal-oriented behaviors.

Figure 6 shows activity of the external cortices during a simulated task lasting 30 min. The picture outlines the first exploratory phase lasting circa 10 min: the DA outflow is increased each time a new cue is perceived and maintenance is entrained by the DA bursts when the correct motor/attention combination is found. The second phase (10–21 min) shows visual exploration and scarce arm activity: the mechatronic board has been widely explored and the cues are no longer able to elicit strong activity in the HIP.

At the beginning of the test phase, a reward becomes visible in one of the boxes. When this is detected, the high release of DA (via Amg) would make the loop maintain the wrong selection (attention on the box and any randomly selected arm-action at the time the reward is perceived): this problem is offset by the fact that the ventral loop is also activated, due to the combined effect of the high DA and the renewed activity in the HIP which also responds to high DA release. Due to the cortico-cortical

connections established during the exploration phase, activity in the PFC plays the role that, in the single loop model, was ascribed to the external input. Provided the weights are strong enough, the PFC activity is then able to bias the selection in both other loops. **Figure 6** shows this process of the PFC biasing the selections each time the reward is moved from one box to another (21, 24, and 27 min): within a few seconds after the reward is perceived, PFC makes the manipulation loop switch to the reach/press channel and the attentional loop switch to the selection of the button associated with the box containing the reward. Attentional loop and input reaching the NAcc are strongly connected (due to the activity in the HIP) so that when the first switches toward the button, the ventral system receives an input related to this new focus. If the correct button is pressed, the focus changes back on the box containing the reward, due to the action of the SC: the input reaching the ventral system changes again and this system eventually restarts biasing the attentional loop to focus on the associated button. This closed causal chain generates an oscillation of both attention and goal between the two targets, i.e., causing a switch of goals from an intermediate one (reach/press the correct button) to the ultimate one (reach the box to secure the food).

When the reward is moved from one box to another, the release of DA decreases, allowing the start of visual exploration until the new position of the reward is detected. Provided the agent has



enough time to explore the whole environment and learn all the associations, it will be then able to solve the task.

3.4. INTRINSIC MOTIVATIONS AND DA CONTROL

Despite the 18 combinations of possible actions on the available targets in the environment (three actions times six cues), the simulated agent usually completes the exploration of all the possible combinations and successfully learns the three associations (after repeating each of them 5–10 times) within the first 10–12 min of a trial. For a comparison, the former version of the model, which exploits a bias in favor of the reach/press action on any target cue (due to the cortico-cortical weights established during learning), took nonetheless an average of 30 min to complete the learning process (Baldassarre et al., 2012).

The behavioral advantage of the new model is evident in the embodied tests carried out on the iCub: the robot is slower than its abstract counterpart and thus needs to maintain its selections until the button is completely pushed to open the box and turn the light on. The benefits accruing from the fact that attention is preserved on a single cue are twofold: first, it allows sufficient amount of time to try several randomly generated actions thereby increasing the chances of selection and completion of a “reach/press”; secondly, it indirectly allows the agent to discriminate between cues that have been already explored and cues that are still novel. By favoring unexplored cues, the agent avoids wasting time trying actions on explored ones and focusses on those that might still allow discovery of novel interactions.

This result arises directly from the manner in which different DA outflows (either caused by intrinsic or extrinsic motivations) alter the agent behavior, narrowing the information provided by the environment. The same mechanism described for the single loop dynamics is replicated in each of the three loops involved in solving the mechatronic board task, but it is triggered by different DA outflows. It is due to this different sensitivity that different effects (e.g., “maintenance” and “weak selection”) may be experienced at the same time in two different striatocortical loops of the same agent. This differentiation allows the agent to fixate attention on novel cues at a certain DA whereas the same agent repeats those actions causing unexpected changes in the environment and fixate on the goal to pursue at a different—higher—DA outflow.

The solution to the behavioral problems arising from the task can be also used to address the problem of the recorded effect DA has on the switch between model free and model-based behaviors (Wunderlich et al., 2012). During the first phase, the agent freely explores the environment guided by its random input, which resembles activity within sensorial cortices reaching the manipulation and attentional loops. This exploration can be considered as “model free” in the restricted sense that the agent does not yet have an explicit model of the environment it is exploring and it is therefore guided by the stimuli in the environment (simulated by noise). On the other hand, the more the process of learning—guided in this task by intrinsically motivating stimuli—allows establishing associations between PFC and both PMC and FEF, the more activity in PFC has the potential to bias the selection in these areas. Thus, when a reward is perceived and the PFC is activated (the ventral loop requires mid-to-high release of DA to be active), its signals guide the whole process of selection performed

in both attention and manipulation loops (**Figure 7**, right), simulating the effect of selections guided by an acquired model of the environment and in particular of the correct combination of action on target (the button) and the resulting effect on a different cue in the environment (the box opens and the light flashes).

4. DISCUSSION

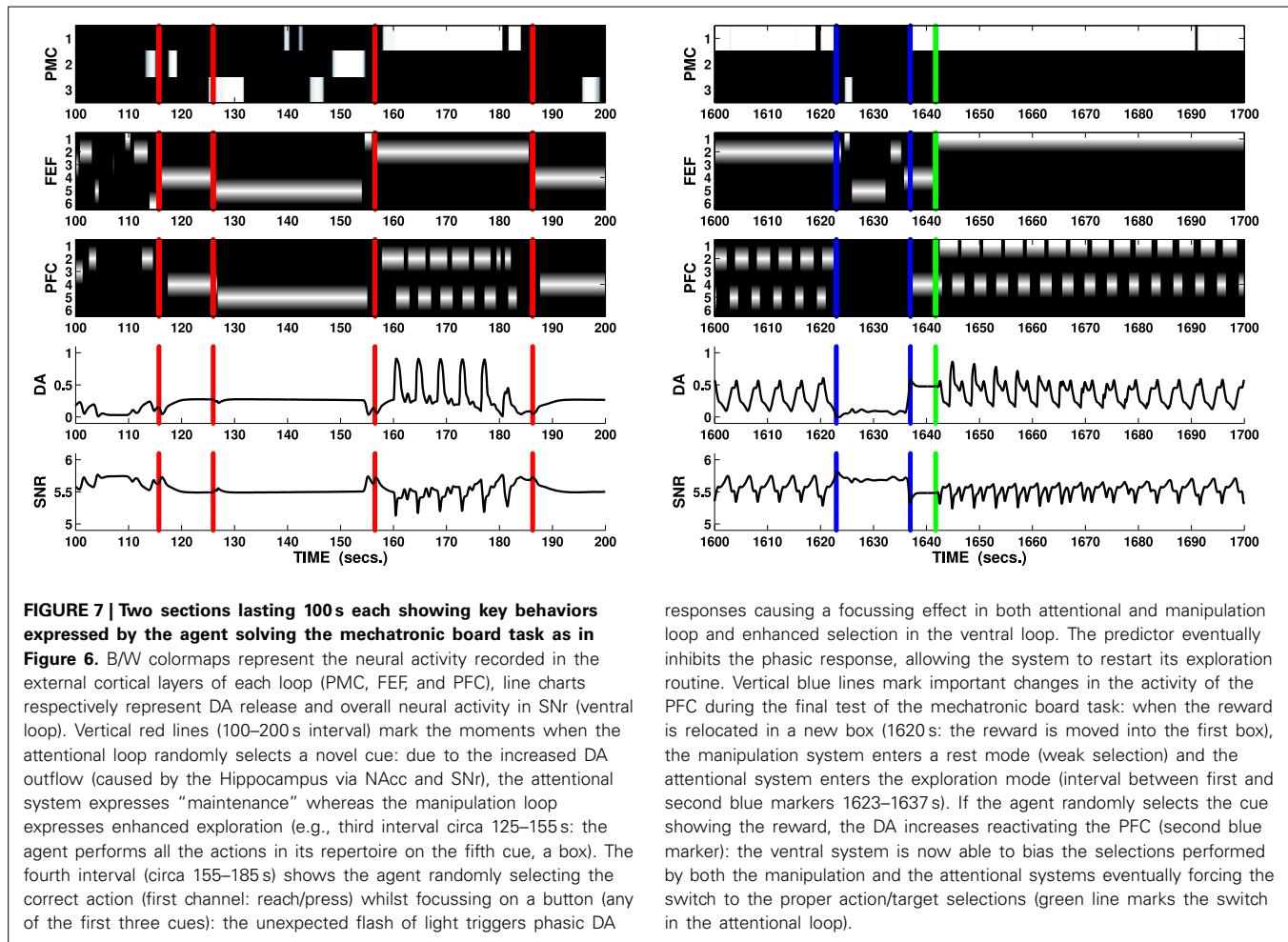
The models we describe show an heterogeneous set of phenomena caused by DA affecting the working status of basal ganglia circuitry: in particular, our tests show a mechanism underlying these phenomena in the dynamic unbalancing of competition established between the direct (via D1 striatum) and hyperdirect (via STN) pathways, with high DA outflows favoring the former.

All the phenomena here simulated and tested on the iCub can be properly considered as emergent: the timing differences bringing forth vigor, maintenance causing the “focus effect” and incentive salience, the dynamic switch between behavioral strategies (rest, exploration, goal oriented behavior and model-based exploitation) do not require *ad hoc* functions or structures to be realized but instead result from intrinsic features characterizing the interaction between DA and Basal Ganglia.

The existence of segregated loops within the circuitry linking cortex and basal ganglia is currently widely accepted when considering macroscopic structures for motor, associative and limbic neural regions (Joel and Weiner, 2000; Kelly and Strick, 2004; Miyachi, 2009): within these macroscopic structures, the exact extent of the channels has been described for motor selection (Alexander and Crutcher, 1990; Romanelli et al., 2005) and the hypothesis that there are similar structures in other macroloops is consistent with findings about segregated values and saliences within the NAcc (Samejima et al., 2005; Lau and Glimcher, 2008; Fitzgerald et al., 2012).

To the best of our knowledge, the models exploiting the functions expressed by this fine grained “channeled circuitry” either investigate the effects of DA on selections performed by feed-forward models of the basal ganglia which do not close the striatocortical loop (Gurney et al., 2001b; Humphries et al., 2006, 2012), or neglect of tonic DA outflow as a regulator of the selections among distinct channels (Prescott et al., 2006; Baldassarre et al., 2012; Chersi et al., 2013).

The former type of models simulate differences in selection strength and distribution that is correlated with different DA outflows (Gurney et al., 2001b; Humphries et al., 2012), but they do not show the accumulation of signal responsible for strong alterations in selection in the presence of low and mid DA outflows. In a similar manner, the difference in the circuitry allows for an explanation of impaired switching only in terms of multiple selections (Humphries et al., 2006) but it cannot simulate the phenomena of interference and maintenance, which are described here as taking place when the DA outflow is still lower than that required for the multiple activations. Humphries et al. (2006, 2012) reach a similar conclusion about DA being responsible for an inverted U effect on the agent’s ability to switch selections following changes in the environment (i.e., the external stimuli), but the differences in the neural architecture allow the present models to provide a more detailed account of the way the input is processed, especially in presence of mid-to-high DA outflows. The



present models not only point out that a successful gating effect is inversely correlated with the mean value/salience ascribed to the input and directly correlated with its sparseness: the models show how, by way of the unbalanced competition, DA outflow eventually affects the spectrum of inputs that can be successfully processed by a striato-cortical loop, either increasing or decreasing it. Furthermore, the presence of the loop allows the system to maintain the performed selections making of each selection a part of the input in the following cycle, ignoring most changes in the environment. From a computational perspective, the use of a loop to create a memory-like phenomenon and preserve neural activity despite changes in the input is not novel: a similar conclusion about preserving selections (there called “latching”) has been reached for instance by Humphries and Gurney (2002). The novelty of the present study is to show how this mechanism can be caused by the dynamics of the DA outflows, hence becoming strongly correlated with the presence of motivations and rewards.

There are concerns that may be raised when establishing a comparison between the biological complexity of the neural structure of the striato-cortical loop and its simplified version implemented in our models. In particular, the results might be biased by three distinct features characterizing the architecture of the present models: first and foremost, the lack of the basal ganglia

indirect pathway; second, the lack of the re-entrant cortico-thalamic loop; finally, the presence of the lateral inhibitions in the second layer of the cortex, which may perform the selections in place of the basal ganglia (as seen in Figure 4, +50% condition).

These concerns may lead to the conclusion that the simulations generated by the models are ill-grounded. However, it should be considered that the effects on selections are mainly due to the alteration of the gain obtained unbalancing the competition between direct and hyperdirect pathways due to increased release of DA, i.e., a condition paralleled and strengthened, in real brain, by a diminished activity in the indirect pathway (due to the presence of D2). Since the indirect pathway plays a major role in regulating the selections by controlling, via GPe, the activity of STN (Gurney et al., 2001a,b; Frank, 2006), the results coming from the model might have a quantitative bias in providing predictions about the amount of DA required to unbalance the competition between direct and hyperdirect pathway, but they should be sufficiently reliable in providing a general qualitative understanding on the consequences of such unbalance.

The lack of the re-entrant thalamo-cortical loop is also a possible cause of biased results. Furthermore, the present model shows a “collapsed” version of the cortical layers involved where the actual biological circuitry (Douglas and Martin, 2004; da Costa

and Martin, 2010) is condensed in a single layer receiving input from the thalamus and propagating it back to the striatum, whilst a second layer is mainly used as an output source for the robotic set-up. Former models (Humphries and Gurney, 2002) have demonstrated the ability of a more complex thalamo-cortical circuitry to preserve a selection in the cortex independently of the input provided by the basal ganglia. Still, the computation performed at this level should not affect our key hypothesis about the role played by the DA in biasing the gain of the striato-cortical loop in favor of the direct pathway. From a computational perspective, the input reaching the striatum from the cortex is weighted by the presence of the DA in the area: as a result, the differences between the single values characterizing each component of this input are increased when the DA outflow increases. After this input is processed in the thalamo-cortical circuit and propagated back to the striatum, this process is repeated, so that the new cycle further increases the differences in the inputs. In the present model the computation performed in the thalamus is simplified via its basal activity, which is lowered by the inhibition provided by the SNr or GPI. A more bio-constrained model would be grounded on reciprocal connections between thalamus and cortex and these would be the cause for the initial activation of the former. We argue that this change might once again affect timing and duration of maintenance, but it would not affect the general hypothesis about the improved gain in the bigger loop involving the striatum, which is essential for the increased chance the basal ganglia have to maintain any performed selection, realizing a memory-like phenomenon. The thalamo-cortical loop is part of the striato-cortical one, so it has for sure an important effect on this maintenance, but the functions of the two structures can be considered as distinct, although affecting each other.

In future work, we plan to model an architecture of the basal ganglia including both the indirect pathway and a more complex thalamo-cortical connectivity, including the reentrant loop, though relying on the same type of computational tools and assumptions. This will allow a better comparison with the known literature via the analysis of how the functioning of this neural system is modulated when the DA release either increases above or decreases below the baseline.

Concerning the selection in the second layer of the cortex, **Figure 4** shows that the mechanism of the lateral inhibition becomes important only when the DA release reaches very high values (e.g., in the single loop test, compare +50% with +20%, +30%, and five out of six intervals in the +40% condition), determining multiple selections in the first layer of the cortex. DA release recorded in most of the task is well below this threshold (see **Figure 7**), so that it is fair to state that the selections performed during the task are properly determined by the basal ganglia rather than by the lateral inhibition in the second layer of the cortex. An example of selections performed without the help of the lateral inhibition is provided by the ventral loop (which lacks these inhibitions in both cortical layers), where it is possible to see some overlap among selections, whilst the system is still able to perform quick switches depending on its input. It is important to stress here that the lack of lateral inhibition in the ventral system is meant mainly for the purpose to demonstrate the ability of the underlying system to perform its selections

independently of the final “filter” which would be implemented in the second cortical layer. This assumption does not entail that the PFC does not have lateral inhibition, as the cortex of the other two loops do. Adding these inhibitions would have resulted in a “cleaner” output signal as the one recorded in the second layers of the attentional and manipulation loop, but it would have possibly concealed the selection of basal ganglia targeted here.

Compared to its early version (Baldassarre et al., 2012), the three-looped model has been modified mainly by altering cortico-cortical connectivity, erasing direct inputs to Cau and Put, adding the hippocampal input to the ventral loop and an agency guided predictor to stop DA bursts when a perceived stimulus is no longer unexpected. What is more important, both DA outflow dynamics and effects it plays in its target regions have been sophisticated. To solve the mechatronic board task, the model exploits the temporary focus effect, jointly with a differentiated sensitivity to DA in different striatal regions. The combination of these two features results in completely different behaviors in relation to distinct DA outflows. It is useful to distinguish three phases in the task: first, the agent visually explores the environment looking for new cues and performing few arm actions; secondly, the agent focusses attention on a new cue and randomly explores possible interactions with the cue itself thanks to its action repertoire; finally, the agent repeats those action selections responsible for generating intrinsically motivating changes in the environment or granting access to rewards.

The early version of the model also had to secure a similar behavior in presence of intrinsic motivations to boost learning processes. To this purpose, a “repetition bias” (Gurney et al., 2009; Baldassarre et al., 2012) was used in the former model. This is a transient process resembling learning and unlearning conceived to offset the well-known (in reinforcement learning field) tendency of a system to stick with the action/procedure it has learned, avoiding any subsequent exploration of the environment (a nice review of the exploration versus exploitation problem can be found in Cohen et al., 2007). This classic problem has been overcome in the present model by simply relying on the differential DA release triggered by either intrinsically (i.e., novel cues and agency-related unexpected changes in the environment) or extrinsically motivating stimuli (i.e., food): we have shown both tonic and phasic DA can be causative in selection maintenance so that even if there were no learning processes, the agent would nonetheless repeat the behavior selected when the motivating stimuli are perceived.

This mechanism, jointly with the effect DA has on accumulation and selection timing, mediates vigor-like behaviors in the agent (see **Figure 5**), suggesting these can be caused by the ability to quickly accumulate signals and preserve a selection rather than by biasing learning processes. The differentiation between repeated behavior and learning denotes a significant difference with respect to classic reinforcement learning models (Niv et al., 2007), but it does not entail these two phenomena do not concur in determining the agent’s overall behavior. It is important to stress that the model described in Baldassarre et al. (2012) had both cortico-striatal and cortico-cortical plasticity. The present model, which aims to investigate more in depth the DA role, does not entail that cortico-striatal learning is not involved in this task.

It rather points out that this learning process, though sufficient for supporting the desired behavioral changes, is not strictly necessary. The removal of such learning allows the current model to better isolate some effects of DA that are often overlooked. In particular, the present model shows that DA, aside its importance for striatal learning, has also a dynamic transient effect on striato-cortical loops, which results in a behavior resembling the one caused by learning. Any learning taking place in the striatum, though biologically plausible due to the presence of DA bursts and surely present in tasks as those considered here, would have made this dynamic effect of DA much less evident, hence was removed from the present model.

On the contrary, the mechatronic board task shows the “focus effect” enhancing both the cortical learning process during exploration and the exploitation after recalling: after the PFC has successfully biased the selections performed in the manipulation and attentional loops, the agent shows a stereotyped behavior pattern in pursuing its reward. In this context DA has still a role in helping the system to focus and maintain its selections, but the learned cortico-cortical connections trigger a switch favoring those selections that are biased by the activity in the ventral loop, rather than those that are temporally close to the increase of DA outflow. The resulting behavior shows both the features described for high vigor (short time reactions) and those characterizing incentive salience (Berridge and Robinson, 1998; Peciña et al., 2003), where the “wanting” is mediated by the stability of the activity of the ventral loop.

Since the functioning realized here is determined by the special features characterizing the neural circuitry of the striatocortical loops, our results show how manipulation, attention and executive control systems may be affected by enhanced selection, interference and maintenance, that in turn are dependent on DA outflow. The model supports the hypothesis that, in normal conditions, different types of motivating stimuli, triggering different DA outflows, modulate selection, but it also provides an interesting explanation of the dysfunctions associated with hyper activation of the D1 receptors mimicking high release of DA in any of the loops. We suggest the so-called “focus effect” in particular may provide a better explanation of the recorded behavior and neural activity associated with intrastriatal injections of amphetamine or DA agonists (Wang and Rebec, 1993; Waszczak et al., 2002; Gulley et al., 2004) or of impulsive/compulsive disorders and stereotyped behaviors in medicated Parkinson’s patients (Weintraub, 2009; Djamshidian et al., 2011).

Data reported in medicated Parkinson’s patients can be explained by the mechanisms we describe in terms of the underlying role of guidance by the ventral loop. The learned cortico-cortical connections represent the acquired associations between actions on specific targets and the resulting changes in the environment, so that when one of these outcomes is desired (i.e., selected in the ventral loop), the learned connections allow the ventral loop to orient the selection in the other systems, causing a switch to a goal-oriented behavior.

The low sensitivity to DA in the ventral loop means this system is only activated in presence of high—tonic or phasic—DA outflows, such as the one caused by either extrinsically or intrinsically motivating stimuli so that it is either active during learning

(establishing the associations) or when exploitation is necessary to pursue a reward (consistently with Wunderlich et al., 2012). But if the agent suffers from a loss of DA release in dorsal striatum and is therefore employing DA agonists to compensate this loss, the ventral striatum (which in Parkinson’s patients is usually less affected by this loss) might be activated much more frequently in contexts which are normally not connected with either extrinsic or intrinsic motivations. The more frequent selections in the ventral loop due to artificially high presence of DA—or even the fixated selection if the DA is sufficiently high—would bias any other selection either in the motor or associative loops and would therefore lead to an artificially induced hyper-incentivized salience on the perceived stimuli and therefore to compulsive behaviors.

Despite the functional analogies that can be established between the motor exploration of a biological system and its artificial simulation presented here, we note that the current implementation of the actions in the robot generates a behavior which, in both conditions of increased DA release, might lead to some misinterpretations. Indeed, both the condition expressing motor exploration of the possible interactions with a novel cue (**Figure 6**, 0–10 min) and the one expressing exploitation of the known associations when either intrinsic or extrinsic rewards is perceived (**Figure 6**, 21–30 min) might resemble a dysfunctional behavior. In particular, during motor exploration several actions are initiated and do not reach their conclusion whereas in presence of motivations the robot expresses a strongly stereotyped behavior. When analyzing these data it is important to remember that the repertoire of actions in this set-up is limited enough to grant a good test (18 possible combinations of actions on different targets) of the effect of DA in narrowing down the options and guiding exploration, but far from being close to the repertoire of actions and environment interactions that would characterize—for instance—a child or a primate when playing with the very same mechatronic board. This is of course a strong limit for the potential variety and flexibility of the final behavior. Furthermore, DA affects the maintenance of a selection performed by a system which is otherwise completely guided (in its attentional and motor selections) by the random walk set in the thalamus. Thus, it is not surprising that the actions are often initiated and then interrupted when DA outflow is not sufficient to perform a strong lock. Using random noise to initiate actions granting the autonomous exploration of an environment is a functional simplification of the real motor exploration performed by biological agents, which is likely guided by goals and constantly affected by the presence of minor intrinsic and extrinsic motivations that can be found in a rich environment. This common procedure in the field of developmental robotics (Saegusa et al., 2009; Gottlieb et al., 2013; Ivaldi et al., 2013; Moulin-Frier and Oudeyer, 2013) is sometimes called “motor babbling” and is used to overcome the need to create an otherwise infeasibly rich environment to motivate exploration.

The model described in this paper can explain a wide range of behaviors under minimal assumptions. Furthermore, it is biologically plausible—being grounded in the neuroanatomy of perceptual and action selection systems. Because the model is formulated in terms of neuronal dynamics that are associated

with specific cortical and subcortical structures, it lends itself to dynamic causal modeling of empirical neuronal responses. For example, it is—in principle—possible to use Equation (1) as a model of hidden neuronal activity associated with sources of electrophysiological responses. By equipping this neuronal model with a conventional electromagnetic forward model, one can then estimate the parameters (connectivity) of the model using non-invasive EEG or MEG measurements. Crucially, one could also evaluate the Bayesian model evidence for dynamic causal models with and without dopaminergic gating or gain control implicit in Equation (1). This would nicely parallel the face validity we have established through implementation of the scheme in a neurorobotics setting.

ACKNOWLEDGMENTS

The Wellcome Trust Centre for Neuroimaging is supported by core funding from the Wellcome Trust (091593/Z/10/Z). This work was supported by the Wellcome Trust (Ray Dolan Senior Investigator Award 098362/Z/12/Z) and the European Commission 7th Framework Programme (FP7/2007–2013), “Challenge 2—Cognitive Systems, Interaction, Robotics,” grant agreement No. ICT-IP-231722, project “IM-CLeVeR—Intrinsically Motivated Cumulative Learning Versatile Robots.”

SUPPLEMENTARY MATERIAL

The compiled file of the C++ libraries, the files containing all the parameters and the essential matlab code to plot graphs here described and any other recording in relation to the present simulations, plus a series of demonstrative graphs showing the neural activity in all layers of the three loops, can be downloaded here: <http://www.im-clever.eu/resources/models/models/fiorewetal2013keepfocussingmodel.tar.gz>

The video showing the iCub accomplishing its task can be found here: <http://www.youtube.com/watch?v=vW6Gf2A3-XQ>

REFERENCES

- Alexander, G. E., and Crutcher, M. D. (1990). Functional architecture of basal ganglia circuits: neural substrates of parallel processing. *Trends Neurosci.* 13, 266–271. doi: 10.1016/0166-2236(90)90107-L
- Alexander, G. E., DeLong, M. R., and Strick, P. L. (1986). Parallel organization of functionally segregated circuits linking basal ganglia and cortex. *Annu. Rev. Neurosci.* 9, 357–381. doi: 10.1146/annurev.ne.09.030186.002041
- Baldassarre, G., Mannella, F., Fiore, V. G., Redgrave, P., Gurney, K., and Mirolli, M. (2012). Intrinsically motivated action-outcome learning and goal-based action recall: a system-level bio-constrained computational model. *Neural Netw.* 41, 168–187. doi: 10.1016/j.neunet.2012.09.015
- Baldassarre, G., and Mirolli, M. (eds.). (2013). *Intrinsically Motivated Learning in Natural and Artificial Systems*. Berlin: Springer-Verlag. doi: 10.1007/978-3-642-32375-1
- Beckstead, R. M., Wooten, G. F., and Trugman, J. M. (1988). Distribution of d1 and d2 dopamine receptors in the basal ganglia of the cat determined by quantitative autoradiography. *J. Comp. Neurol.* 268, 131–145. doi: 10.1002/cne.902680113
- Beierholm, U., Guitart-Masip, M., Economides, M., Chowdhury, R., Düzel, E., Dolan, R., et al. (2013). Dopamine modulates reward related vigor. *Neuropsychopharmacology* 38, 1495–503. doi: 10.1038/npp.2013.48
- Berridge, K. C., and Robinson, T. E. (1998). What is the role of dopamine in reward: hedonic impact, reward learning, or incentive salience? *Brain Res. Brain Res. Rev.* 28, 309–369. doi: 10.1016/S0165-0173(98)00019-8
- Cabib, S., and Puglisi-Allegra, S. (2012). The mesoaccumbens dopamine in coping with stress. *Neurosci. Biobehav. Rev.* 36, 79–89. doi: 10.1016/j.neubiorev.2011.04.012
- Chersi, F., Mirolli, M., Pezzulo, G., and Baldassarre, G. (2013). A spiking neuron model of the cortico-basal ganglia circuits for goal-directed and habitual action learning. *Neural Netw.* 41, 212–224. doi: 10.1016/j.neunet.2012.11.009 (Special Issue on Autonomous Learning).
- Chevalier, G., and Deniau, J. M. (1990). Disinhibition as a basic process in the expression of striatal functions. *Trends Neurosci.* 13, 277–280. doi: 10.1016/0166-2236(90)90109-N
- Cohen, J. D., McClure, S. M., and Yu, A. J. (2007). Should I stay or should I go? how the human brain manages the trade-off between exploitation and exploration. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 362, 933–942. doi: 10.1098/rstb.2007.2098
- Cools, R., and D’Esposito, M. (2011). Inverted-u-shaped dopamine actions on human working memory and cognitive control. *Biol. Psychiatry* 69, e113–e125. doi: 10.1016/j.biopsych.2011.03.028
- da Costa, N. M., and Martin, K. A. C. (2010). Whose cortical column would that be? *Front. Neuroanat.* 4:16. doi: 10.3389/fnana.2010.00016
- Dayan, P., and Abbott, L. (2005). *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*. Cambridge, MA: MIT Press.
- Djamshidian, A., Cardoso, F., Grosset, D., Bowden-Jones, H., and Lees, A. J. (2011). Pathological gambling in parkinson’s disease—a review of the literature. *Mov. Disord.* 26, 1976–1984. doi: 10.1002/mds.23821
- Douglas, R. J., and Martin, K. A. C. (2004). Neuronal circuits of the neocortex. *Annu. Rev. Neurosci.* 27, 419–451. doi: 10.1146/annurev.neuro.27.070203.144152
- Durstewitz, D. (2009). “Neurocomputational analysis of dopamine function,” in *Dopamine Handbook*, eds L. Iversen, S. Iversen, S. Dunnett, and A. Bjorklund (Oxford: Oxford University Press), 261–276.
- Düzel, E., Bunzeck, N., Guitart-Masip, M., and Düzel, S. (2010). Novelty-related motivation of anticipation and exploration by dopamine (nomad): implications for healthy aging. *Neurosci. Biobehav. Rev.* 34, 660–669. doi: 10.1016/j.neubiorev.2009.08.006
- Fellous, J. M., and Linster, C. (1998). Computational models of neuromodulation. *Neural Comput.* 10, 771–805. doi: 10.1162/089976698300017476
- FitzGerald, T. H. B., Friston, K. J., and Dolan, R. J. (2012). Action-specific value signals in reward-related regions of the human brain. *J. Neurosci.* 32, 16417–16234. doi: 10.1523/JNEUROSCI.3254-12.2012
- Floresco, S. B., West, A. R., Ash, B., Moore, H., and Grace, A. A. (2003). Afferent modulation of dopamine neuron firing differentially regulates tonic and phasic dopamine transmission. *Nat. Neurosci.* 6, 968–973. doi: 10.1038/nn1103
- Frank, M. J. (2006). Hold your horses: a dynamic computational role for the subthalamic nucleus in decision making. *Neural Netw.* 19, 1120–1136. doi: 10.1016/j.neunet.2006.03.006
- Frank, M. J. (2011). Computational models of motivated action selection in cortico-striatal circuits. *Curr. Opin. Neurobiol.* 21, 381–386. doi: 10.1016/j.conb.2011.02.013
- Frank, M. J., Samanta, J., Moustafa, A. A., and Sherman, S. J. (2007). Hold your horses: impulsivity, deep brain stimulation, and medication in parkinsonism. *Science* 318, 1309–1312. doi: 10.1126/science.1146157
- Frank, M. J., Seeberger, L. C., and O’reilly, R. C. (2004). By carrot or by stick: cognitive reinforcement learning in parkinsonism. *Science* 306, 1940–1943. doi: 10.1126/science.1102941
- Friston, K. J., Shiner, T., FitzGerald, T., Galea, J. M., Adams, R., Brown, H., et al. (2012). Dopamine, affordance and active inference. *PLoS Comput. Biol.* 8:e1002327. doi: 10.1371/journal.pcbi.1002327
- Friston, K. J., Tononi, G., Reike, G. N., Sporns, O., and Edelman, G. M. (1994). Value-dependent selection in the brain: simulation in a synthetic neural model. *Neuroscience* 59, 229–243. doi: 10.1016/0306-4522(94)90592-4
- Gottlieb, J., Oudeyer, P.-Y., Lopes, M., and Baranes, A. (2013). Information-seeking, curiosity, and attention: computational and neural mechanisms. *Trends Cogn. Sci.* 17, 585–593. doi: 10.1016/j.tics.2013.09.001
- Grace, A. A., Floresco, S. B., Goto, Y., and Lodge, D. J. (2007). Regulation of firing of dopaminergic neurons and control of goal-directed behaviors. *Trends Neurosci.* 30, 220–227. doi: 10.1016/j.tins.2007.03.003
- Grillner, S., Hellgren, J., Ménard, A., Saitoh, K., and Wikström, M. A. (2005). Mechanisms for selection of basic motor programs—roles for the striatum and pallidum. *Trends Neurosci.* 28, 364–370. doi: 10.1016/j.tins.2005.05.004
- Guitart-Masip, M., Huys, Q. J. M., Fuentemilla, L., Dayan, P., Düzel, E., and Dolan, R. J. (2012). Go and no-go learning in reward and punishment: interactions between affect and effect. *Neuroimage* 62, 154–166. doi: 10.1016/j.neuroimage.2012.04.024

- Gulley, J. M., Reed, J. L., Kuwajima, M., and Rebec, G. V. (2004). Amphetamine-induced behavioral activation is associated with variable changes in basal ganglia output neurons recorded from awake, behaving rats. *Brain Res.* 1012, 108–118. doi: 10.1016/j.brainres.2004.03.044
- Gurney, K., Humphries, M., and Redgrave, P. (2009). Cortico-striatal plasticity for action-outcome learning using spike timing dependent eligibility. *BMC Neurosci.* 10:E135. doi: 10.1186/1471-2202-10-S1-P135
- Gurney, K., Lepora, N., Shah, A., Koene, A., and Redgrave, P. (2013). “Action discovery and intrinsic motivation: a biologically constrained formalisation,” in *Intrinsically Motivated Learning in Natural and Artificial Systems*, eds G. Baldassarre and M. Mirolli (Berlin: Springer-Verlag).
- Gurney, K., Prescott, T. J., and Redgrave, P. (2001a). A computational model of action selection in the basal ganglia. I. A new functional anatomy. *Biol. Cybern.* 84, 401–410. doi: 10.1007/PL00007984
- Gurney, K., Prescott, T. J., and Redgrave, P. (2001b). A computational model of action selection in the basal ganglia. II. Analysis and simulation of behaviour. *Biol. Cybern.* 84, 411–423. doi: 10.1007/PL00007985
- Haber, S. N., Fudge, J. L., and McFarland, N. R. (2000). Striatonigrostriatal pathways in primates form an ascending spiral from the shell to the dorsolateral striatum. *J. Neurosci.* 20, 2369–2382.
- Hikosaka, O. (2007). Gabaergic output of the basal ganglia. *Prog. Brain Res.* 160, 209–226. doi: 10.1016/S0079-6123(06)60012-5
- Hikosaka, O., Sesack, S. R., Lecourtier, L., and Shepard, P. D. (2008). Habenula: crossroad between the basal ganglia and the limbic system. *J. Neurosci.* 28, 11825–11829. doi: 10.1523/JNEUROSCI.3463-08.2008
- Hikosaka, O., Takikawa, Y., and Kawagoe, R. (2000). Role of the basal ganglia in the control of purposive saccadic eye movements. *Physiol. Rev.* 80, 953–978.
- Hotson, J. R., Langston, E. B., and Langston, J. W. (1986). Saccade responses to dopamine in human MPTP-induced parkinsonism. *Ann. Neurol.* 20, 456–463. doi: 10.1002/ana.410200404
- Humphries, M. D., and Gurney, K. N. (2002). The role of intra-thalamic and thalamocortical circuits in action selection. *Network* 13, 131–156. doi: 10.1088/0954-898X/13/1/305
- Humphries, M. D., Khamassi, M., and Gurney, K. (2012). Dopaminergic control of the exploration-exploitation trade-off via the basal ganglia. *Front. Neurosci.* 6:9. doi: 10.3389/fnins.2012.00009
- Humphries, M. D., Stewart, R. D., and Gurney, K. N. (2006). A physiologically plausible model of action selection and oscillatory activity in the basal ganglia. *J. Neurosci.* 26, 12921–12942. doi: 10.1523/JNEUROSCI.3486-06.2006
- Ivaldi, S., Nguyen, M., Lyubova, N., Droniou, A., Padois, V., Filliat, D., et al. (2013). “Object learning through active exploration,” in *IEEE Transactions on Autonomous Mental Development*. doi: 10.1109/TAMD.2013.2280614
- Joel, D., and Weiner, I. (2000). The connections of the dopaminergic system with the striatum in rats and primates: an analysis with respect to the functional and compartmental organization of the striatum. *Neuroscience* 96, 451–474. doi: 10.1016/S0306-4522(99)00575-8
- Kakade, S., and Dayan, P. (2002). Acquisition and extinction in autoshaping. *Psychol. Rev.* 109, 533–544. doi: 10.1037/0033-295X.109.3.533
- Kaplan, F., and Oudeyer, P.-Y. (2007). In search of the neural circuits of intrinsic motivation. *Front. Neurosci.* 1:225–236. doi: 10.3389/neuro.01.1.1.017.2007
- Kelly, R. M., and Strick, P. L. (2004). Macro-architecture of basal ganglia loops with the cerebral cortex: use of rabies virus to reveal multisynaptic circuits. *Prog. Brain Res.* 143, 449–459. doi: 10.1016/S0079-6123(03)43042-2
- Kimchi, E. Y., and Laubach, M. (2009). Dynamic encoding of action selection by the medial striatum. *J. Neurosci.* 29, 3148–3159. doi: 10.1523/JNEUROSCI.5206-08.2009
- Lau, B., and Glimcher, P. W. (2008). Value representations in the primate striatum during matching behavior. *Neuron* 58, 451–463. doi: 10.1016/j.neuron.2008.02.021
- Lisman, J. E., and Grace, A. A. (2005). The hippocampal-vta loop: controlling the entry of information into long-term memory. *Neuron* 46, 703–713. doi: 10.1016/j.neuron.2005.05.002
- Metta, G., Natale, L., Nori, F., Sandini, G., Vernon, D., Fadiga, L., et al. (2010). The icub humanoid robot: an open-systems platform for research in cognitive development. *Neural Netw.* 23, 1125–1134. doi: 10.1016/j.neunet.2010.08.010
- Mink, J. W. (1996). The basal ganglia: focused selection and inhibition of competing motor programs. *Prog. Neurobiol.* 50, 381–425. doi: 10.1016/S0301-0082(96)00042-1
- Mirolli, M., Santucci, V. G., and Baldassarre, G. (2013). Phasic dopamine as a prediction error of intrinsic and extrinsic reinforcements driving both action acquisition and reward maximization: a simulated robotic study. *Neural Netw.* 39, 40–51. doi: 10.1016/j.neunet.2012.12.012
- Miyachi, S. (2009). Cortico-basal ganglia circuits—parallel closed loops and convergent/divergent connections. *Brain Nerve* 61, 351–359.
- Montague, P. R., Hyman, S. E., and Cohen, J. D. (2004). Computational roles for dopamine in behavioural control. *Nature* 431, 760–767. doi: 10.1038/nature03015
- Moulin-Frier, C., and Oudeyer, P.-Y. (2013). “Exploration strategies in developmental robotics: a unified probabilistic framework,” in *IEEE International Conference on Development and Learning and Epigenetic Robotics, IEEE ICDL-Epirob* (Osaka), 1–6. doi: 10.1109/DevLrn.2013.6652535
- Nakano, K. (2000). Neural circuits and topographic organization of the basal ganglia and related regions. *Brain Dev.* 22, S5–S16. doi: 10.1016/S0387-7604(00)00139-X
- Niv, Y., Daw, N. D., Joel, D., and Dayan, P. (2007). Tonic dopamine: opportunity costs and the control of response vigor. *Psychopharmacology (Berl.)* 191, 507–520. doi: 10.1007/s00213-006-0502-4
- Peciña, S., Cagniard, B., Berridge, K. C., Aldridge, J. W., and Zhuang, X. (2003). Hyperdopaminergic mutant mice have higher “wanting” but not “liking” for sweet rewards. *J. Neurosci.* 23, 9395–9402.
- Piggott, M. A., Marshall, E. F., Thomas, N., Lloyd, S., Court, J. A., Jaros, E., et al. (1999). Dopaminergic activities in the human striatum: rostrocaudal gradients of uptake sites and of d1 and d2 but not of d3 receptor binding or dopamine. *Neuroscience* 90, 433–445. doi: 10.1016/S0306-4522(98)00465-5
- Prescott, T. J., González, F. M. M., Gurney, K., Humphries, M. D., and Redgrave, P. (2006). A robot model of the basal ganglia: behavior and intrinsic processing. *Neural Netw.* 19, 31–61. doi: 10.1016/j.neunet.2005.06.049
- Ranganath, C., and Rainer, G. (2003). Neural mechanisms for detecting and remembering novel events. *Nat. Rev. Neurosci.* 4, 193–202. doi: 10.1038/nrn1052
- Redgrave, P., and Gurney, K. (2006). The short-latency dopamine signal: a role in discovering novel actions? *Nat. Rev. Neurosci.* 7, 967–975. doi: 10.1038/nrn2022
- Redgrave, P., Gurney, K., and Reynolds, J. (2008). What is reinforced by phasic dopamine signals? *Brain Res. Rev.* 58, 322–339. doi: 10.1016/j.brainresrev.2007.10.007
- Redgrave, P., Prescott, T. J., and Gurney, K. (1999a). The basal ganglia: a vertebrate solution to the selection problem? *Neuroscience* 89, 1009–1023. doi: 10.1016/S0306-4522(98)00319-4
- Redgrave, P., Prescott, T. J., and Gurney, K. (1999b). Is the short-latency dopamine response too short to signal reward error? *Trends Neurosci.* 22, 146–151. doi: 10.1016/S0166-2236(98)01373-3
- Redgrave, P., Rodriguez, M., Smith, Y., Rodriguez-Oroz, M. C., Lehericy, S., Bergman, H., et al. (2010). Goal-directed and habitual control in the basal ganglia: implications for parkinson’s disease. *Nat. Rev. Neurosci.* 11, 760–772. doi: 10.1038/nrn2915
- Reed, P., Mitchell, C., and Nokes, T. (1996). Intrinsic reinforcing properties of putatively neutral stimuli in an instrumental two-lever discrimination task. *Anim. Learn. Behav.* 24, 38–45. doi: 10.3758/BF03198952
- Romanelli, P., Esposito, V., Schaal, D. W., and Heit, G. (2005). Somatotopy in the basal ganglia: experimental and clinical evidence for segregated sensorimotor channels. *Brain Res. Brain Res. Rev.* 48, 112–128. doi: 10.1016/j.brainresrev.2004.09.008
- Ryan, R. M., and Deci, E. L. (2000). Intrinsic and extrinsic motivations: classic definitions and new directions. *Contemp. Edu. Psychol.* 25, 54–67. doi: 10.1006/ceps.1999.1020
- Saezuga, R., Metta, G., Sandini, G., and Sakka, S. (2009). “Active motor babbling for sensorimotor learning,” in *IEEE International Conference on Robotics and Biomimetics* (Bangkok), 794–799. doi: 10.1109/ROBIO.2009.4913101
- Salamone, J. D., and Correa, M. (2012). The mysterious motivational functions of mesolimbic dopamine. *Neuron* 76, 470–485. doi: 10.1016/j.neuron.2012.10.021
- Salamone, J. D., Correa, M., Mingote, S., and Weber, S. M. (2003). Nucleus accumbens dopamine and the regulation of effort in food-seeking behavior: implications for studies of natural motivation, psychiatry, and drug abuse. *J. Pharmacol. Exp. Ther.* 305, 1–8. doi: 10.1124/jpet.102.035063
- Samejima, K., Ueda, Y., Doya, K., and Kimura, M. (2005). Representation of action-specific reward values in the striatum. *Science* 310, 1337–1340. doi: 10.1126/science.1115270

- Schultz, W. (2006). Behavioral theories and the neurophysiology of reward. *Annu. Rev. Psychol.* 57, 87–115. doi: 10.1146/annurev.psych.56.091103.070229
- Schultz, W. (2007). Multiple dopamine functions at different time courses. *Annu. Rev. Neurosci.* 30, 259–288. doi: 10.1146/annurev.neuro.28.061604.135722
- Schultz, W., Romo, R., Scarnati, E., Sundström, E., Jonsson, G., and Studer, A. (1989). Saccadic reaction times, eye-arm coordination and spontaneous eye movements in normal and MPTP-treated monkeys. *Exp. Brain Res.* 78, 253–267. doi: 10.1007/BF00228897
- Surmeier, D. J., Ding, J., Day, M., Wang, Z., and Shen, W. (2007). D1 and d2 dopamine-receptor modulation of striatal glutamatergic signaling in striatal medium spiny neurons. *Trends Neurosci.* 30, 228–235. doi: 10.1016/j.tins.2007.03.008
- Sutton, R. S., and Barto, A. G. (1998). *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press.
- Taffoni, F., Formica, D., Schiavone, G., Scorcia, M., Tomassetti, A., di Sorrentino, E. P., et al. (2013). “The “mechatronic board”: a tool to study intrinsic motivations in humans, monkeys, and humanoid robots,” in *Intrinsically Motivated Learning in Natural and Artificial Systems*, eds G. Baldassarre and M. Mirolli (Berlin: Springer), 411–432. doi: 10.1007/978-3-642-32375-1_16
- Tishby, N., and Polani, D. (2011). “Information theory of decisions and actions,” in *Perception-Action Cycle. Springer series in cognitive and neural systems*, eds V. Cuturidis, A. Hussain, and J. G. Taylor (New York, NY: Springer), 601–636.
- Utter, A. A., and Basso, M. A. (2008). The basal ganglia: an overview of circuits and function. *Neurosci. Biobehav. Rev.* 32, 333–342. doi: 10.1016/j.neubiorev.2006.11.003
- Wang, Z., and Rebec, G. V. (1993). Neuronal and behavioral correlates of intrastriatal infusions of amphetamine in freely moving rats. *Brain Res.* 627, 79–88. doi: 10.1016/0006-8993(93)90751-8
- Waszczak, B. L., Martin, L. P., Finlay, H. E., Zahr, N., and Stellar, J. R. (2002). Effects of individual and concurrent stimulation of striatal d1 and d2 dopamine receptors on electrophysiological and behavioral output from rat basal ganglia. *J. Pharmacol. Exp. Ther.* 300, 850–861. doi: 10.1124/jpet.300.3.850
- Weintraub, D. (2009). Impulse control disorders in parkinson’s disease: prevalence and possible risk factors. *Parkinsonism Relat. Disord.* 15, S110–S113. doi: 10.1016/S1353-8020(09)70794-1
- Wunderlich, K., Smittenaar, P., and Dolan, R. J. (2012). Dopamine enhances model-based over model-free choice behavior. *Neuron* 75, 418–424. doi: 10.1016/j.neuron.2012.03.042
- Znamenskiy, P., and Zador, A. M. (2013). Corticostriatal neurons in auditory cortex drive decisions during auditory discrimination. *Nature* 497, 482–485. doi: 10.1038/nature12077

Conflict of Interest Statement: The Review Editor, Dimitris Pinotsis, declares that, despite being affiliated to the same institution as authors Vincenzo G. Fiore, Karl Friston and Raymond J. Dolan, the review process was handled objectively and no conflict of interest exists. The Review Editor, Jennifer Lewis, declares that, despite being affiliated to the same institution as author Kevin Gurney, the review process was handled objectively and no conflict of interest exists. The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 05 July 2013; accepted: 29 January 2014; published online: 21 February 2014.

Citation: Fiore VG, Sperati V, Mannella F, Mirolli M, Gurney K, Friston K, Dolan RJ and Baldassarre G (2014) Keep focussing: striatal dopamine multiple functions resolved in a single mechanism tested in a simulated humanoid robot. *Front. Psychol.* 5:124. doi: 10.3389/fpsyg.2014.00124

This article was submitted to Cognitive Science, a section of the journal *Frontiers in Psychology*.

Copyright © 2014 Fiore, Sperati, Mannella, Mirolli, Gurney, Friston, Dolan and Baldassarre. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



No learning where to go without first knowing where you're coming from: action discovery is trajectory, not endpoint based

Martin Thirkettle^{1,2*}, Thomas Walton², Peter Redgrave², Kevin Gurney² and Tom Stafford²

¹ Department of Psychology, The Open University, Buckinghamshire, UK

² Department of Psychology, University of Sheffield, Sheffield, UK

Edited by:

Marco Mirolli, Istituto di Scienze e Tecnologie della Cognizione, Italy

Reviewed by:

Matthew Schlesinger, Southern Illinois University, USA

John Spencer, University of Iowa, USA

***Correspondence:**

Martin Thirkettle, Department of Psychology, The Open University, Walton Hall, Milton Keynes, Buckinghamshire MK7 6AA, UK
e-mail: martin.thirkettle@open.ac.uk

Intrinsic motivations drive an agent to explore, providing essential data for linking behaviors with novel outcomes and so laying the foundation for future flexible action. We present experiments using a new behavioral task which allows us to interrogate the connection between exploration and action learning. Human participants used a joystick to search repeatedly for a target location, only receiving feedback on successful discovery. Feedback delay was manipulated, as was the starting position. Experiment 1 employed stable starting positions, so the task could be learnt with respect to a target location or a target trajectory. Participants were able to learn the correct movement under all delay conditions. Experiment 2 used a variable starting location, so the correct movement could only be learnt in terms of target location. Participants displayed little to no learning in this experiment. These results suggest that movements on this scale are stored as trajectories rather than in terms of target location. Overall the experiments demonstrate the potential of this task for uncovering the native representational substrates of action learning.

Keywords: action outcome learning, spatial learning, intrinsic motivation, cognitive science, motor learning

INTRODUCTION

When Thorndike (1911) pioneered the study of action acquisition with his puzzle box escape paradigm, he was investigating whether animals can learn to produce apparently insightful behavior despite having no causal understanding of the problem at hand. By repeatedly placing subjects into a puzzle box and measuring the time taken for them to enact their escape, Thorndike was able to observe and record the animals as they gradually extracted the elements of behavior associated with success from a complex stream of self-generated behavioral variance. Thorndike's animals improved across trials, and while they may have had little insight into the underlying relationship between their actions and escape, the feat of learning was nonetheless impressive, requiring them to solve a considerable problem of credit assignment (Minsky, 1961; Sutton and Barto, 1998). The major challenge of learning through trial and error is that success will inevitably be associated with both causally relevant and causally irrelevant activities. In addition to this, the learning system must deal with delays between successful actions and their associated outcomes—the so called “distal reward problem” (Izhikevich, 2007), with no way of determining how far back in the motor record the most important aspects of performance might lie.

By associating motor activity with a particular outcome, animals create an action-outcome pair, which can then be added to the behavioral repertoire. Theories of the representation of action suggest that it is the outcome which is represented after learning (Hommel et al., 2001), but these focus on the goal of the action, rather than how to perform the action. During

action discovery, especially in a situation where discovery occurs during unconstrained exploration, the contingency is not necessarily obvious, and moreover identifying the causal element of motor output is non-trivial. In normative models of reinforcement learning, a common method is to use temporal difference algorithms which maintain a trace of the pattern of recent activity, such that it remains eligible for reinforcement at the moment when the outcome eventually occurs (Barto et al., 1981; Wickens, 1990; Singh and Sutton, 1996). This approach is consistent with Skinner's studies of superstitious behavior (1948), and predicts that participants will learn sub-optimal strategies based on prior success, as previously successful trajectories of movement will be reinforced regardless of the underlying contingency of the action-outcome pair. It is also clear that this mechanism of associating recent motor activity with success leaves little opportunity for insight into which aspect of the previously successful movement is causal and which can be pruned across repetitions. Such refinements could only occur through a process of trial-and-error across numerous action repetitions.

Redgrave and Gurney (2006) and Redgrave et al. (2008) have argued that the response of dopamine neurons in the ventral midbrain ~100 ms after the presentation of novel and rewarding stimuli acts as an indiscriminate time-stamp which indicates the last segment of the animal's motor record that could have played a role in eliciting a novel stimulus, irrespective of what that stimulus might be. They propose that the dopamine response is central to the tasks of agency detection, action discovery and the learning of action-effect contingencies. It is widely thought that this activity plays a key role in

valuation and economic decision-making (Schultz et al., 1997; Schultz, 2007), and in the case of action discovery, Redgrave and Gurney (2006) suggest that the dopamine response acts as the timestamp against which the motor commands in the eligibility trace can be compared—ameliorating the distal reward problem. While this time-stamping mechanism prevents any motor commands subsequent to the outcome, and therefore non-contingent, from entering the pool of potential contingencies, the record of recent motor output eligible for reinforcement will still contain non-contingent elements, and any manipulation of the time between movement performance and success being signaled will necessarily introduce further non-contingent motor output.

The twin problems of credit assignment and distal reward are at the heart of a paradigm we created to investigate this kind of “Thorndikian” action acquisition in human and animal participants (Stafford et al., 2012). This task captures the discovery of a novel action through self-generated behavior and allows the refinement of that behavior through the trial-and-error pruning of non-contingent elements to be studied. In this task participants move a joystick freely and “escape” the trial by discovering the action set by the experimenter, in this case simply placing the joystick controlled cursor within a pre-defined area. The participants receive no feedback on the cursor’s location and successful performance of this action is denoted only by an audio signal and the end of the trial. Other work using this paradigm has focused on the neural pathways preferred for processing the reinforcing signal (Thirkettle et al., 2013), or on the time sensitivity of these mechanisms (Walton et al., 2013). Together this work seeks to better understand the cognitive mechanisms and neural pathways involved in the discovery and learning of novel actions through self-generated exploratory behaviors. Stafford et al. (2012) include an in depth discussion of the nature of the joystick task and its relationship to previous behavioral work studying learning. The present experiments aim to identify if an “eligibility trace” of movement trajectories generated during an iterated location finding task is necessary for learning. Previous studies have focused on the discovery of a new action-outcome pair; here that moment of discovery is studied alongside the refinement of the action through repetition. If participants learn a novel action by stamping in recent motor output, there should be evidence of this in the form of the preservation of portions of successful movements from early performances in later ones. The design of the joystick task, lacking as it does, any visual information regarding either the target location, or the current location of the joystick in the search arena encourages the participants to use motoric and bodily sources of information. The type of location information used in a location finding task has been shown to affect performance in terms of both systematic biases and absolute levels of performance (e.g., Simmering et al., 2008), but here our focus is on maintaining a constant source of information—proprioception and efference copy—and manipulating the usefulness of relevance of past experience to inform learning. If a reliance on the movements made previously is found, we would predict this would preclude learning in a situation where only the endpoint of a previous movement, rather than the movement itself was informative.

We therefore sought to measure learning performance when the eligibility trace was contaminated with additional, non-contingent, motor commands, and when the record of motor commands was devalued across movement repetitions. In experiment one, participants discovered the location of a hidden target area and then repeated moving to this target from the same start position. By manipulating the delay between the participant entering the target area and the presentation of the success signal, contamination was introduced into the record of recent motor commands. If the eligibility trace is bound by a time-stamping mechanism then increasing this delay between action performance and reinforcement should produce weaker learning and more variable movement trajectories across repetitions. In experiment two, we repeated the manipulation of delay in the location finding task but used a randomized starting location for each repetition of the movement to the target—forcing participants to return to the target area from a different position each time. If participants are relying on the previous movements rather a representation of the target location to refine their performance across repetitions then both learning and performance should be extremely poor.

EXPERIMENT 1

MATERIALS AND METHODS

Participants

Thirty undergraduate students (mean age 19 years) at the University of Sheffield (25 females) participated in all conditions of this study. Participants took part in return for credits in the department’s research participation scheme. All subjects were naive to the purpose of the experiment and the independent variable. Ethical approval was granted by the department’s ethics committee.

Apparatus

The experimental program was written in Matlab (Version 2007), and stimulus display was performed using the Psychophysics Toolbox extensions (Brainard, 1997). A 19" monitor was used throughout along with a standard USB keyboard for participant response during instructions. A commercial joystick (Logitech extreme 3D pro joystick, P/N: 863225-1000) was used as the experimental input device. Custom Matlab code polled the position of the joystick at 100 Hz.

The search space was defined as a square with a side length of 1024 units. Movements of the joystick were mapped onto movements within the search space in a one to one fashion, with the joystick starting in the center of the search space at the beginning of each trial. Once released from the grip of a participant, the joystick’s internal spring returned it to the center of the search space within a tolerance of 10 units.

Procedure

Participants sat at a desk in front of the joystick and monitor. Before starting the experimental program, the task was briefly described verbally with the goal being phrased as “finding the correct position to place the joystick in.” Participants were told that the experiment involved no deception and that the correct

position could always be found. Following this brief verbal reassurance, the program was started and the participants were asked to follow the onscreen instructions.

The size of the target area (hotspot) that participants were required to find was determined through pilot testing and set to occupy 0.28% of the search space. Experimental trials were split into ten iterations, an initial iteration where they had to search for the hotspot, and nine subsequent iterations where they had to return to a hotspot in the same position. Each iteration began with the joystick in the center of the search space and for each trial the center of the hotspot was positioned randomly within an annulus shaped region of the total search space to ensure that the hotspot never overlapped the central starting point or the outer edge of the search space. During an iteration any movement of the joystick into this region was defined as a “hit” and resulted in a beep (600 Hz) of 10 ms duration and the end of the iteration. During each iteration, the screen remained dark and blank. A delay period of 0, 150, 300, or 450 ms was interposed between the moment at which a participant moved into the hotspot and the point at which the beep was presented. This also marked the end of the current iteration and was accompanied by an on-screen message to prepare for the next (see Appendix 1 for full details of onscreen text).

Before the experimental trials, participants completed a short practice session and once this was completed the experimental trials began immediately. Participants completed 2 trials at each delay duration, each trial containing 10 iterations—for a total of 80 movements. Order of trial delay condition was counterbalanced across participants to control for order effects.

Data analysis

We used a 4th order two-way low pass Butterworth filter at 10 Hz to remove noise and redundant data points from the movement data. This filter is commonly used in studies of human motion (Seidler, 2007) and smoothed the raw joystick output, the intention being to more accurately reflect the underlying movement of the participant’s hand and arm.

For the purposes of analysis, the trace of movement from each iteration was treated as being composed of two phases: pre-discovery and post-discovery (Figure 1). The pre-discovery period extends from the start of the iteration to the point of entry into the hotspot and is free to last as long as the participant takes to discover the target. Conversely, the duration of the post-discovery period is strictly dictated by the delay imposed by the experimenter between the successful discovery of the target and the presentation of success signal. The post-discovery period is of particular interest as it contains contaminating—non-contingent—information produced by the participant. If people learn by stamping in recent motor output, their activity during this period should influence their performance during the pre-discovery period of subsequent trials.

Due to the open-ended nature of trials, it was anticipated prior to testing that the distribution of trial duration and distance covered would be positively skewed. Analysis of the data distributions confirmed this and all data were corrected using log-transformation prior to analysis (Keene, 1995).

RESULTS

In each iteration, the movement within the pre-discovery period can be compared against a direct line from the start position to the target. We term the difference between this straight line and the path taken by the participant the “irrelevant distance” and by collapsing this measure across the ten iterations of each trial, across trials and across participants the impact of the imposed signal delays on performance before contact was made with the target can be assessed.

As Figure 2 shows, there was a significant effect of delay on pre-discovery irrelevant distance, $F_{(3, 87)} = 4.79, p < 0.005$, driven entirely by the slump in performance in the 450 ms condition ($p < 0.05$). This is a consistent, albeit less dramatic, effect to that reported in our previous work using a manipulation of signal delay in the joystick task, and we attribute the reduced delay sensitivity found here compared to the previous work to the iterated nature of the present task.

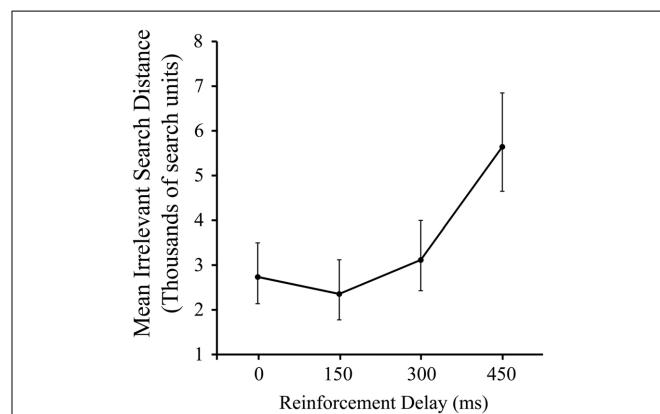
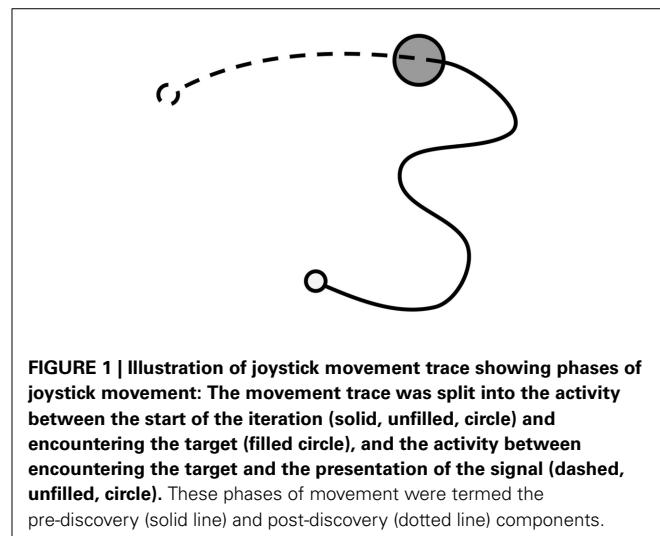


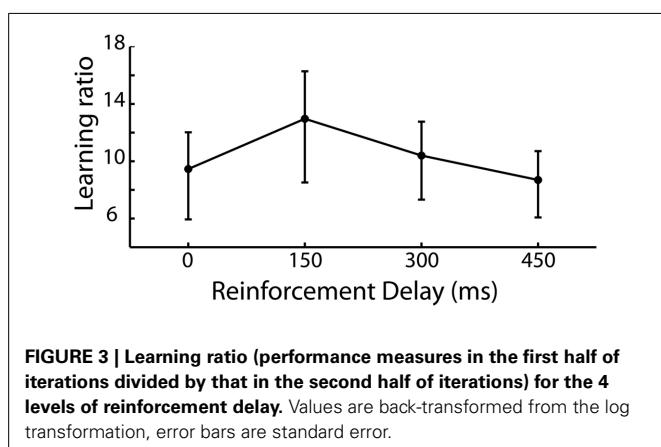
FIGURE 2 | Mean irrelevant pre-discovery distance (and standard error) for the 4 levels of reinforcement delay. Values are back-transformed from the log transformation.

The requirement of the participant to repeat a newly acquired action allows the learning of that action to be captured and as such we expected to see a reduction in the irrelevant distance moved by the participant in later iterations of trials compared to earlier iterations of the same trial. A learning ratio was calculated by dividing the irrelevant distance traveled in the pre-discovery period in iterations 1–5 by that traveled in iterations 6–10 of each trial. **Figure 3** shows the learning ratio collapsed across participants for the four delay conditions and shows that while there was no significant effect of delay on learning [$F_{(3, 87)} = 0.142, p = 0.935$], a considerable improvement in performance was observed from early to late trials. This suggests that the effect of delay on learning impacts action discovery, rather than the refinement of the discovered action across the later iterations.

Within each trial, the refinement of the newly discovered action across iteration can be approximated by the power law of learning (Ritter and Schooler, 2001) as in the equation:

$$\text{efficiency} = Em + \text{range} \times e^{-\alpha N}$$

Here we use irrelevant search distance as the measure of efficiency, with α being the parameter which describes the speed

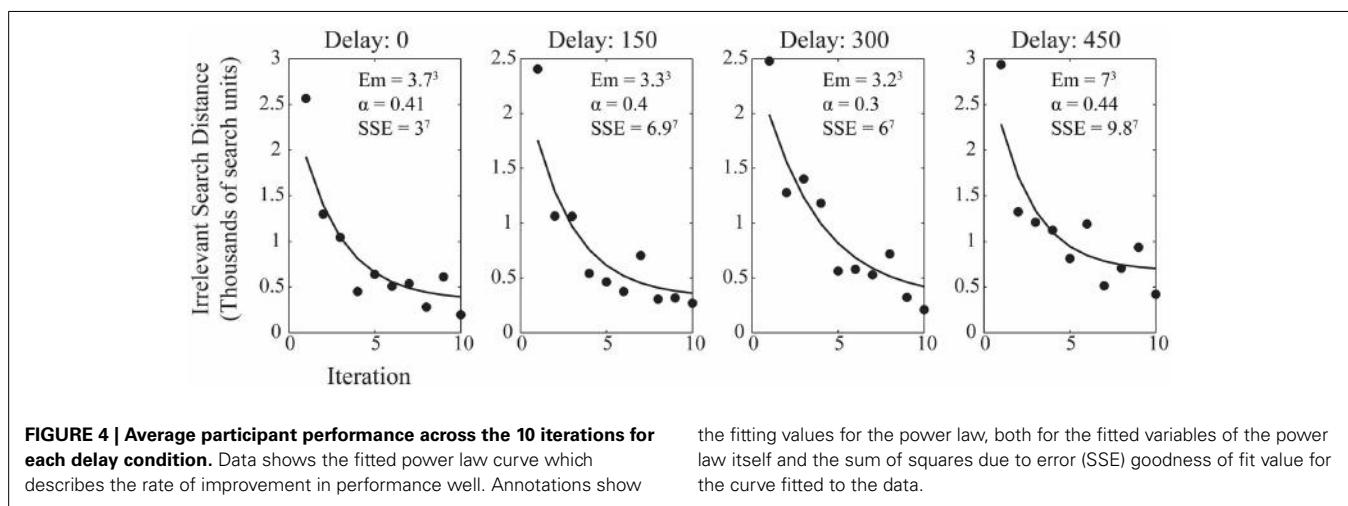


of learning with the range of observed performance levels, Em is a minimum figure for the irrelevant search distance, range is the difference between initial and asymptotic performance, and N is the number of trials. **Figure 4** shows the average performance of participants within each trial at each reinforcement delay condition with the best power law fit applied. The improvement in performance is well-described by the power law at each level of delay, although, again, it is notable that the greater delay had more of an impact on the minimum irrelevant search distance than upon the value of α which describes the rate at which performance improved to asymptote. The similarity of α across delay conditions, as with the learning ratio, suggests that delay is impeding action discovery, rather than action refinement.

The learning evident in experiment 1, shows that after discovering the invisible target location in the first iteration, the movement required by the task was refined across the subsequent 9 iterations. Participants were able to greatly reduce the length of their path to the target by the final iteration compared to that taken on their first encounter with the target.

EXPERIMENT 2

A limitation of using a stable starting position when seeking to investigate whether a particular trajectory of movement is stamped into behavior is that if participants perform close to optimally in an early trial, it is difficult to determine on subsequent trials whether similar trajectories of movement are a reflection of the participants adopting previously successful trajectories of movement or whether they are adopting previously successful trajectories simply because these trajectories are consistent with near optimal performance and they would have learnt to perform at this level anyway. An alternative approach is to vary the start position whilst maintaining a stable target. In this way it is possible to ensure that the optimal trajectory of movement varies from trial to trial, making it easier to determine whether participants are exploiting their memory of a previously successful movement path or learning a successful end point which they are able to reach from any starting



position. Experiment 2 therefore replicated experiment 1, but after discovering the target location instead of repeating the movement from the same start position to the target 9 times, participants moved to the target from 9 randomly chosen start positions.

MATERIALS AND METHODS

Participants

15 undergraduate students (mean age 19 years) at the University of Sheffield (11 females) participated in all conditions of this study. Again, participants took part in return for credits in the department's research participation scheme. All subjects were naive to the purpose of the experiment and the independent variable.

Apparatus

The experimental setup remained as in experiment 1; with the exception that stimulus display was performed using the Cambridge Research Systems Visage graphics board and the associated Matlab toolbox extensions. A Mitsubishi Diamond Pro 2070sb 22" monitor was used throughout and a chin rest ensured the participants remained seated 57 cm from the screen throughout. Changes to the experimental code meant that the position of the joystick was now polled at 1000 Hz and the search space was defined as a square with a side length of 1000 units.

Procedure

All experimental procedures were kept as similar as possible as in experiment 1, with the addition of a requirement that the participant moved the joystick to a randomly selected start position before beginning the search on each iteration for the target. Also three rather than four different delay levels were employed in order to reduce experiment time, and focus more tightly on the most influential delays between success and reinforcement signal presentation. The randomization of start location was achieved by presenting the start position and the current position of the joystick on-screen and instructing the participant to move the cursor to the highlighted area in order to start the iteration. The start position was chosen in the same way as the target position (which, as in experiment 1, remained unchanged for the 10 iterations of each trial), with the additional constraint that it could not overlap the target position. As in experiment 1 participants understood that the target position was changing only for each trial of 10 iterations.

Participants again completed a short practice session immediately before the experimental trials and conducted three trials of 10 iterations at each of three delay levels (0, 200, and 400 ms) for a total of 90 trials. The resulting data was processed in the same way as in experiment 1 to correct for positive skew, and reduce redundant data points from the movement data.

RESULTS

There was a significant effect of delay on the irrelevant distance traveled by the participants when the start position was randomized [$F_{(2, 28)} = 13.422, p < 0.001$]. As in experiment 1, this effect was driven entirely by the highest level

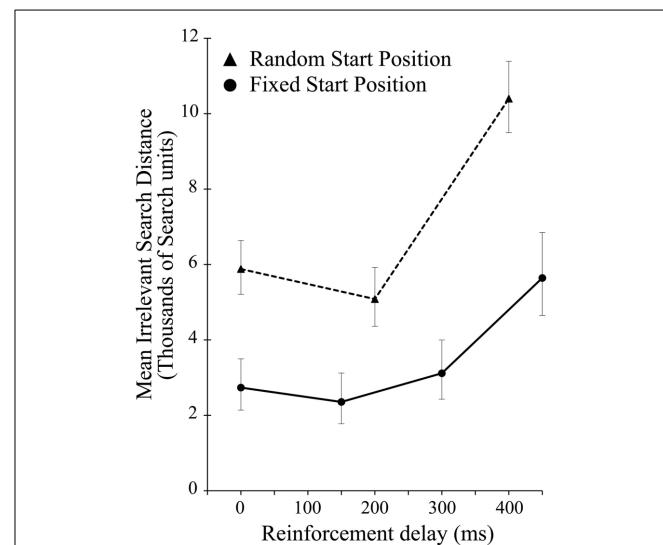


FIGURE 5 | Mean irrelevant pre-discovery distance (and standard error) for the 3 levels of reinforcement delay in experiment 2 (shown as dotted line). The results from experiment 1 (solid line) are also plotted for comparison.

of delay, in this case 400 ms [$F_{(1, 14)} = 16.290, p = 0.001$]. As Figure 5 shows, across all delay conditions, comparing between the experiments participants average irrelevant distance was greater when seeking a static target if the start position was changed from iteration to iteration suggesting a reliance on the static start position in order to find the unchanging target. While the delay manipulation is unequal across the two experiments, preventing in depth analysis, comparing performance in just the zero delay conditions, shows that changing the start position significantly impaired performance [$t_{(40,135)} 2.709, p = 0.01$].

A simple increase in the irrelevant distance traveled could signify that by changing the start position on each iteration the task of finding the target was made more difficult, rather than speaking to the effect of the changing start position on learning. However, this is revealed in the measures of learning across the ten iteration of experiment 2. Figure 6 shows the learning ratio and fitted power law data for performance in experiment 2. Again, we see no significant effect of delay on the learning ratio but the lack of improvement across the 10 iterations is striking. Unlike in experiment 1, participants did not improve as the repetitions of the movement continued, and this is borne out in a significant reduction in the learning ratio. Comparing the zero delay conditions across the two experiments again we see this reduction is significant [$t_{(35,724)} 5.776, p < 0.01$]. The lack of learning found without a stable start position is evidenced further by our attempts to fit a power law to the data as in experiment 1. The power law of learning no longer describes the data as no improvement in performance of any note is taking place. This strongly suggests that the refinement of the newly discovered action as found in experiment 1 is heavily reliant on a stable start position.

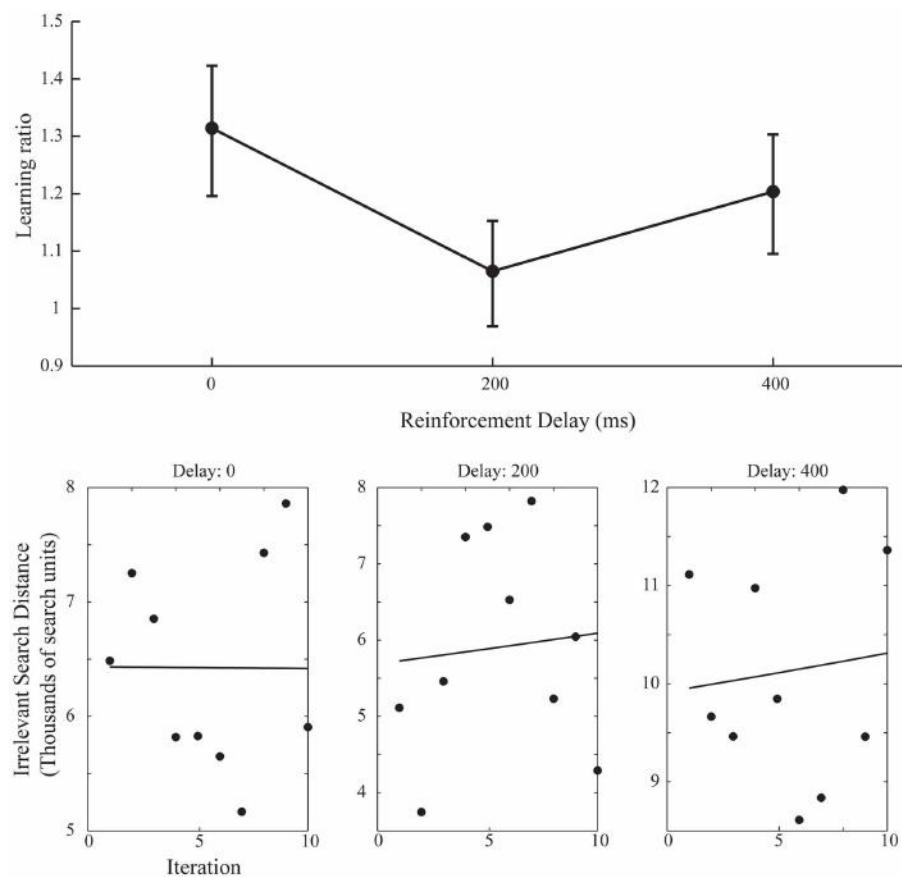


FIGURE 6 | Learning ratios and fitted performance data for experiment 2. Compared to the learning ratios of experiment 1 (Figure 3) we again see no effect of delay on learning ratio but dramatically reduced values for said ratios (upper axis). Unlike in experiment 1 (Figure 4) the power law curve shown on the lower

set of axis no longer adequately fits the data and participants did not reliably improve across the 10 iterations (Power law curve fit values for lower axis figures: 0ms Delay: $E_m = 3.7^3$, $\alpha = 5.2^{-4}$, SSE = 7.1^6 , 200ms Delay: $E_m = 1.6^3$, $\alpha = -9.4^{-3}$, SSE = 1.7^7 , 400ms Delay: $E_m = 6.6^3$, $\alpha = -1.1^{-2}$, SSE = 1.2^7).

GENERAL DISCUSSION

Whether the starting position was static (experiment 1) or changed with each iteration (experiment 2) participants were able to successfully discover the target location, but are affected by delaying the reinforcement signal. This is consistent with our previous findings using different versions of this task (Stafford et al., 2012; Thirkettle et al., 2013; Walton et al., 2013), but here we demonstrate the impact of reinforcement delay in a version of the task in which the reinforcement is delivered without giving the participant the opportunity to correct for, or respond to, the delay within a single performance of the reinforced action. Here participants had to repeat the entire action after a single, possibly delayed, reinforcement signal. This sensitivity to delay reveals, we argue that it is the process of action acquisition which is critically dependent on the coincidence of motor efference copy with a sensory signal indicating a novel or surprising outcome (Redgrave and Gurney, 2006) rather than the subsequent refinement of a discovered action through, in this case, repeated encounters with the target area. For the initial discovery of an action, the number of potentially causative movements grows with each moment and this record inevitably becomes increasingly contaminated

with noise (i.e., movements or aspects of movement with no causative relationship to the action). Because of this, delivery of the sensory signal to the brain area(s) where it can be used to tag potentially causal elements in the motor record must be done as fast as possible in order to reduce the difficulty of the credit assignment problem (Minsky, 1961). In machine learning, the idea of an “eligibility trace” has been suggested as a mechanism for solving the credit assignment problem (Singh and Sutton, 1996). With regard to the joystick task, a system employing such an “eligibility trace” should display the repetition of aspects of movement contained within such a period regardless of their necessity for success. Further studies are planned to focus on the production and persistence of these “superstitious movements” in the joystick task.

Learning to move to a spatial target is significantly poorer, indeed, almost abolished, when only the target location remains static and the participant must move to the target without reference to their previous movements (experiment 2). This allows an additional supposition about how the credit assignment problem is being solved here: Not only is learning in this task achieved by a highly time sensitive mechanism, such as an eligibility

trace', but that this mechanism operates on a record of previous movements not a record of previous locations. The task in experiment 1 could be solved by a learning mechanism that stored target information, or trajectory information (since learning to move toward the target location, or moving in a target direction would both allow successful completion of the task). The target location method remains viable for experiment 2, but the task cannot be successfully completed by acquiring target trajectories—the start location of the movement shifts, requiring different trajectories to reach the target. The absence of learning in experiment 2, but successful learning in experiment 1, suggests that the participants are relying on a trajectory based strategy.

That a stable starting position could be so critical to learning is somewhat surprising. Previous work with an emphasis on spatial goals has shown that both rats and humans are capable of learning even when a stable trajectory is not associated with the goal (Tolman, 1948; Landau et al., 1984). Human visuo-spatial reasoning is highly developed and, for example, in tasks such as the pursuit rotor task (Frith and Lang, 1979) participants are able to trace a moving target so that current spatial position guides trajectory. In the Morris Water Maze (Morris, 1984) rats learn a target location rather than a trajectory or by using "dead reckoning" [but see Chamizo (2003)]. That our participants are not able to use spatial location to guide their movements suggests that our task taps a different set of processes. Indeed, we designed the task (Stafford et al., 2012) to rely as far as possible on the processes of motor learning without augmentation from visual-spatial memory or explicit reasoning. By using a task that tapped implicit motor processes we hoped to be able to isolate the specific capacities of this architecture of action discovery. Alternatively, it is also possible that the lack of reliable visuo-spatial information in this formulation of the task forces the system to rely on trajectory information to an unusual extent. Certainly the addition of spatial information in the form of a visual cue would have colored the results, and further experimentation is required to assess the relative contribution of each category of information on learning. However, what can be said with certainty is that in the absence of visuo-spatial information the system is capable of using only trajectory information from efference copy to learn spatial tasks.

If we consider how an animal might learn under natural conditions, it seems likely that the behavior it chooses to reselect—in effect, the unit of reinforcement—might relate to the attitude of its body and its overall position within the environment at the moment when reinforcement arrived. The ability to learn particular trajectories of movement might not be a key aspect of action acquisition because reinforcement is so rarely contingent on such movement. Indeed, there is mounting evidence to suggest that the motor output an animal is most inclined to reselect and reinforce might be its terminal body position rather than the movement trajectory required to achieve that body position. Graziano (2006) describes how attempts to map the motor cortex have revealed that actions do not appear to be represented at the neural level in the form of motor primitives that can be combined to form complete actions. Instead, particular portions of the cortex, when stimulated, evoke whole meaningful adaptive

responses such as defensive or feeding postures. Furthermore, certain aspects of these actions appear to be more important than others. For example, hand movements are encoded in such a way that the hand will finish at a specific point in space, irrespective of where it started. Such representations do not describe a particular sequence of movements and instead describe behaviorally relevant terminal postures. In Graziano's view, certain features of actions, such as the final hand position, are crucial and the means by which these positions are achieved are of less importance and are likely free to vary to a greater extent. These representations in the cortical behavioral repertoire are plastic, and are able to represent complex movements as a function of experience and training (Martin et al., 2005; Ramanathan et al., 2006). While our current results may appear in tension with this body of evidence, one reconciliation is that regardless of the final representation in the cortex (which seemingly does include the terminal posture), a stable trajectory of movement is sufficient to support this process of learning. In other words, the conditions required for learning actions can be different from the eventual form of their storage.

These experiments validate the task as being a useful one for investigating the mechanisms of novel action learning (Stafford et al., 2012). The manipulation of delay allows us to expose the time sensitivity of these mechanisms (Walton et al., 2013), while precise stimulus control even allows us to discern the involvement of different neural pathways in action learning (Thirkettle et al., 2013). The current result suggests that trajectories can act as the substrate of novel action learning and further that in the absence of both visual information and a stable trajectory, actions can be discovered but cannot be refined over subsequent repetitions, although it should be noted that it remains possible that with more repetitions some improvement in performance could be observed.

We were inspired in this investigation by our theory of the function of the basal ganglia in novel action learning (Redgrave and Gurney, 2006; Redgrave et al., 2013). These variations of the "joystick task" are important and revealing as a whole because action acquisition presents a particularly difficult problem in the compromise between over-constrained and under-constrained tasks: when we over-constrain, we leave little opportunity for the agent to generate interesting behavioral variance as they freely explore and discover the new action; but when we under-constrain, there is simply too much noise in the data for us to draw any meaningful conclusions.

This work has been inspired by considering human action learning from the perspective of an autonomous agent which must acquire novel actions without either explicit instruction or certain knowledge of action-outcome relations (see also Shah et al., submitted). We have been guided in this by work in intrinsically motivated learning, and particularly by work within the framework of reinforcement learning (Sutton and Barto, 1998). From this reinforcement learning perspective a number of direct predictions flow. For example, the exploration-exploitation dilemma is a fundamental trade-off in learning within a complex space of actions where the reinforcement signal has unknown bounds. Early focus on actions with highest known value may lead to failure to discover the highest value actions in the long run, and—conversely—early exploration may lead to the discovery of

the highest value actions in the long-term. In the joystick task this predicts that those participants who “explore” the movement space more in early trials—i.e., those who cover a greater distance reaching the target—will eventually settle on a more optimal path than those who explore less and “exploit” a sufficient path to the target. We have confirmed that this signature of an exploration-exploitation trade-off manifests in our joystick task (Stafford et al., 2012) as well as in at least one other domain of skill acquisition (Stafford and Dewar, 2013).

For an autonomous agent the credit assignment problem is deeply under-constrained—any aspect of the agent’s behavior could potentially be causative of some novel outcome. The present experiment shows that the action learning system of human subjects have a bias to attribute cause to trajectory aspects of brief motor actions, rather than spatial aspects (resulting in the failure to learn seen in experiment 2). It is plausible to suggest that a “representational bias” may exist in these systems in order to narrow down their search of motor space for novel action-outcome pairs. An animal analog of the joystick task has been developed, and research already conducted demonstrates that general measures of behavior are comparable between rat and human participants (Stafford et al., 2012). Further work is required to assess whether the core challenge of the credit assignment problem is approached in a similar manner

REFERENCES

- Baldassarre, G., and Mirolli, M. (eds.). (2013). *Intrinsically Motivated Learning in Natural and Artificial Systems*. New York, NY: Springer. doi: 10.1007/978-3-642-32375-1
- Barto, A. G., Sutton, R. S., and Brouwer, P. S. (1981). Associative search network: a reinforcement learning associative memory. *Biol. Cybern.* 40, 201–211. doi: 10.1007/BF00453370
- Brainard, D. H. (1997). The psychophysics toolbox. *Spat. Vis.* 10, 433–436. doi: 10.1163/156856897X00357
- Chamizo, V. (2003). Acquisition of knowledge about spatial location: assessing the generality of the mechanism of learning. *Q. J. Exp. Psychol. B*, 56, 102–113. doi: 10.1080/02724990244000205
- Frith, C. D., and Lang, R. J. (1979). Learning and reminiscence as a function of target predictability in a two-dimensional tracking task. *Q. J. Exp. Psychol.* 31, 103–109. doi: 10.1080/14640747908400710
- Graziano, M. (2006). The organization of behavioral repertoire in motor cortex. *Annu. Rev. Neurosci.* 29, 105–134. doi: 10.1146/annurev.neuro.29.051605.112924
- Hommel, B., Müseler, J., Aschersleben, G., and Prinz, W. (2001). The Theory of Event Coding (TEC): A framework for perception and action planning. *Behav. Brain Sci.* 24, 849–878. doi: 10.1017/S0140525X0100103
- Izhikevich, E. M. (2007). Solving the distal reward problem through linkage of STDP and dopamine signaling. *Cereb. Cortex* 17, 2443–2452. doi: 10.1093/cercor/bhl152
- Keene, O. N. (1995). The log transformation is special. *Stat. Med.* 14, 811–819. doi: 10.1002/sim.4780140810
- Landau, B., Spelke, E., and Gleitman, H. (1984). Spatial knowledge in a young blind child. *Cognition* 16, 225–260. doi: 10.1016/0010-0277(84)90029-5
- Martin, J. H., Engber, D., and Meng, Z. (2005). Effect of forelimb use on postnatal development of the forelimb motor representation in primary motor cortex of the cat. *J. Neurophysiol.* 93, 2822–2831. doi: 10.1152/jn.01060.2004
- Minsky, M. (1961). Steps toward artificial intelligence, *Proc. IRE*, 49, 8–30. doi: 10.1109/JRPROC.1961.287775
- Morris, R. (1984). Developments of a water-maze procedure for studying spatial learning in the rat. *J. Neurosci. Methods* 11, 47–60. doi: 10.1016/0165-0270(84)90007-4
- Ramanathan, D., Conner, J. M., and Tuszyński, M. H. (2006). A form of motor cortical plasticity that correlates with recovery of function after brain injury. *Proc. Natl. Acad. Sci. U.S.A.* 103, 11370–11375. doi: 10.1073/pnas.0601065103
- Redgrave, P., and Gurney, K. (2006). The short-latency dopamine signal: a role in discovering novel actions? *Nat. Rev. Neurosci.* 7, 967–975. doi: 10.1038/nrn2022
- Redgrave, P., Gurney, K., and Reynolds, J. (2008). What is reinforced by phasic dopamine signals? *Brain Res. Rev.* 58, 322–339. doi: 10.1016/j.brainresrev.2007.10.007
- Redgrave, P., Gurney, K., Stafford, T., Thirkettle, M., and Lewis, J. (2013). “The role of the basal ganglia in discovering novel actions,” in *Intrinsically Motivated Learning in Natural and Artificial Systems* eds G. Baldassarre and M. Mirolli (New York, NY: Springer), 129–150.
- Ritter, F. E., and Schooler, L. J. (2001). “The learning curve,” *International Encyclopedia of the Social and Behavioral Sciences*, eds W. Kintch, N. Smelser, and P. Baltes (Oxford: Elsevier Ltd), 8602–8605. doi: 10.1016/B0-08-043076-7/01480-7
- Schultz, W. (2007). Multiple dopamine functions at different time courses. *Annu. Rev. Neurosci.* 30, 259–288. doi: 10.1146/annurev.neuro.28.061604.135722
- Schultz, W., Dayan, P., and Montague, P. R. (1997). A neural substrate of prediction and reward. *Science* 275, 1593–1599. doi: 10.1126/science.275.5306.1593
- Seidler, R. D. (2007). Older adults can learn to learn new motor skills. *Behav. Brain Res.* 183, 118–122. doi: 10.1016/j.bbr.2007.05.024
- Simmerling, V. R., Peterson, C., Darling, W., and Spencer, J. P. (2008). Location memory biases reveal the challenges of coordinating visual and kinesthetic reference frames. *Exp. Brain Res.* 184, 165–178. doi: 10.1007/s00221-007-1089-7
- Singh, S. P., and Sutton, R. S. (1996). Reinforcement learning with replacing eligibility traces. *Mach. Learn.* 22, 123–158. doi: 10.1007/BF00114726
- Skinner, B. F. (1948). Superstition in the pigeon. *J. Exp. Psychol.* 38, 168–172. doi: 10.1037/h0055873
- Stafford, T., and Dewar, M. (2013). “Testing theories of skill learning using a very large sample of online game players,” in *22nd Annual Conference of the Cognitive Science Society*.
- Stafford, T., Thirkettle, M., Walton, T., Vautrelle, N., Hetherington, L., Port, M., et al. (2012). A novel task for the investigation of action acquisition. *PloS ONE* 7:e37749. doi: 10.1371/journal.pone.0037749
- Sutton, R. S., and Barto, A. G. (1998). *Reinforcement Learning: An introduction*. Cambridge: Cambridge University Press.

in rats, and whether search strategies and persistent elements of movement are detectable in rat response data and comparable to human data. Should this research reveal a common solution to the credit assignment problem, we would suggest that the use of a representational bias to focus the search of motor space could be a reasonable approach for any artificial agent.

The ability to acquire new actions is a keystone of human intelligence and the drive to explore our motor competency an essential element of our intrinsic motivations, relating, as it does, to aspects of intrinsic motivation such as novelty, surprise, curiosity, and mastery (Baldassarre and Mirolli, 2013). Our task allows us to frame general questions about intrinsic motivation and action discovery within a tightly controlled experimental context.

ACKNOWLEDGMENTS

The research leading to the results presented here was supported by the European Community’s Seventh Framework Programme FP7/2007–2013, “Challenge 2 – Cognitive Systems, Interaction, Robotics,” under grant agreement No. FP7-ICT-IP-231722, project “IM-CLeVeR – Intrinsically Motivated Cumulative Learning Versatile Robots.” Thomas Walton was supported by an EPSRC DTC grant.

- Thirkettle, M., Walton, T., Shah, A., Gurney, K., Redgrave, P., and Stafford, T. (2013). The path to learning: action acquisition is impaired when visual reinforcement signals must first access cortex. *Behav. Brain Res.* 243, 267–272. doi: 10.1016/j.bbr.2013.01.023
- Thorndike, E. L. (1911). *Animal Intelligence: Experimental Studies*. New York, NY: Macmillan. doi: 10.5962/bhl.title.55072
- Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychol. Rev.* 55, 189. doi: 10.1037/h0061626
- Walton, T., Thirkettle, M., Redgrave, P., Gurney, K. N., and Stafford, T. (2013). The discovery of novel actions is affected by very brief reinforcement delays and reinforcement modality. *J. Mot. Behav.* 45, 351–360. doi: 10.1080/00222895.2013.806108
- Wickens, J. (1990). Striatal dopamine in motor activation and reward-mediated learning: steps towards a unifying model. *J. Neural Trans. Gen. Sect.* 80, 9–31. doi: 10.1007/BF01245020

Conflict of Interest Statement: The authors declare that the research

was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 17 May 2013; paper pending published: 27 June 2013; accepted: 28 August 2013; published online: 18 September 2013.

*Citation: Thirkettle M, Walton T, Redgrave P, Gurney K and Stafford T (2013) No learning where to go without first knowing where you're coming from: action discovery is trajectory, not endpoint based. *Front. Psychol.* 4:638. doi: 10.3389/fpsyg.2013.00638*

This article was submitted to Cognitive Science, a section of the journal Frontiers in Psychology.

Copyright © 2013 Thirkettle, Walton, Redgrave, Gurney and Stafford. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Robust active binocular vision through intrinsically motivated learning

Luca Lonini^{1*}, Sébastien Forestier^{1,2}, Céline Teulière¹, Yu Zhao³, Bertram E. Shi³ and Jochen Triesch¹

¹ Frankfurt Institute for Advanced Studies, Goethe University, Frankfurt am Main, Germany

² École Normale Supérieure Cachan Bretagne, Bruz, France

³ Department of Electronic and Computer Engineering, HK University of Science and Technology, Hong Kong, China

Edited by:

Gianluca Baldassarre, Italian National Research Council, Italy

Reviewed by:

Mototaka Suzuki, Humboldt University, Germany

Kathryn E. Merrick, University of New South Wales, Australia

***Correspondence:**

Luca Lonini, Frankfurt Institute for Advanced Studies, Goethe University, Ruth-Moufang Str. 1, 60438 Frankfurt am Main, Germany
e-mail: lonini@fias.uni-frankfurt.de

The efficient coding hypothesis posits that sensory systems of animals strive to encode sensory signals efficiently by taking into account the redundancies in them. This principle has been very successful in explaining response properties of visual sensory neurons as adaptations to the statistics of natural images. Recently, we have begun to extend the efficient coding hypothesis to active perception through a form of intrinsically motivated learning: a sensory model learns an efficient code for the sensory signals while a reinforcement learner generates movements of the sense organs to improve the encoding of the signals. To this end, it receives an intrinsically generated reinforcement signal indicating how well the sensory model encodes the data. This approach has been tested in the context of binocular vision, leading to the autonomous development of disparity tuning and vergence control. Here we systematically investigate the robustness of the new approach in the context of a binocular vision system implemented on a robot. Robustness is an important aspect that reflects the ability of the system to deal with unmodeled disturbances or events, such as insults to the system that displace the stereo cameras. To demonstrate the robustness of our method and its ability to self-calibrate, we introduce various perturbations and test if and how the system recovers from them. We find that (1) the system can fully recover from a perturbation that can be compensated through the system's motor degrees of freedom, (2) performance degrades gracefully if the system cannot use its motor degrees of freedom to compensate for the perturbation, and (3) recovery from a perturbation is improved if both the sensory encoding and the behavior policy can adapt to the perturbation. Overall, this work demonstrates that our intrinsically motivated learning approach for efficient coding in active perception gives rise to a self-calibrating perceptual system of high robustness.

Keywords: active perception, sparse coding, reinforcement learning, robotics, stereo vision, vergence, robustness

1. INTRODUCTION

A number of studies in the last four decades addressed the question of how sensory neurons encode information and showed that neural systems might employ an efficient code to represent incoming data, i.e., a code that exploits redundant information (Attnave, 1954; Barlow, 1961; Field, 1994). The visual system has been a primary target of these studies, where the main result showed that neurons in primary visual cortex (V1) might encode visual information through a *sparse code*, i.e., a code where, at any given moment, only a few neurons out of the entire population fire. A sparse coding strategy has several benefits (Willshaw et al., 1969; Lennie, 2003), including increased memory, less interference between stored patterns and reduced energy consumption, as compared to a dense code (i.e., where many units are simultaneously active). Importantly, when the sparse coding principle is applied to the encoding of natural images (i.e., scenes from nature), it leads to the emergence of basis functions whose structure resemble that of V1 simple cells' receptive fields (Olshausen et al., 1996). The idea of sparse coding has been confirmed by neurophysiological experiments, showing sparse activation of V1

neurons in primates when probed with image sequences of natural stimuli (Weliky et al., 2003) and has been extended to other sensory domains, including the olfactory and auditory domain (Perez-Orive et al., 2002; Smith and Lewicki, 2006). Most studies treated the problem of efficient coding without considering the effects of behavior. The connection between sensory inputs and behavior, commonly referred to as the *perception-action cycle* is important both to (1) understand the development of sensory representations in neural systems as a function of the task performed (Rothkopf et al., 2009) and to (2) design artificial systems, such as robots that autonomously learn and adapt to a changing environment. Indeed, a big technological challenge for such systems is to learn in an efficient and unsupervised way.

We consider this problem in the context of binocular vision. Binocular disparity, the difference between the image projected on left and right retina, is used by organisms with two frontal eyes as a primary depth cue. In order to focus on a point at a certain depth, the two eyes are required to jointly turn inwards or outwards, such that the same object or world feature appears in the center of both images and disparity is nullified. Such type of

eye movement is known as *vergence* and represents a fundamental component of visually-guided behavior.

Many approaches to perform vergence in robotic systems employ computer vision techniques to estimate disparity from stereo-images followed by the use of a feedback controller to move the eyes and nullify disparity. These methods are often dependent on pre-defined system parameters and camera calibration. Some methods have used reinforcement learning to autonomously learn vergence control; however, they all require estimating disparity by the use of a pre-defined set of filters (Piater et al., 1999) or a population of disparity-selective neurons (Franz and Triesch, 2007; Wang and Shi, 2010).

In our previous work (Zhao et al., 2012; Lonini et al., 2013c) we have presented a method that autonomously learns how to verge two cameras on a common world feature based on the efficient coding hypothesis. The model makes use of a form of *intrinsic motivation* to learn efficient sensory representations in the perception-action cycle. A sparse coding model learns to encode sensory information using binocular basis functions at different resolutions, while a reinforcement learner generates the camera movement, according to the output of the sparse coding model. Sensory coding and behavior develop in parallel, by minimizing the same cost function: the error between the original stimulus and its reconstruction by the sparse coding model. The rationale behind the approach is that, the more similar left and right images are, the easier they are to encode. Thus, if the actions taken by the reinforcement learning (RL) agent drive the system to perform correct vergence, the reconstruction error will be minimized. Importantly, the reward to the reinforcement learning agent is generated within the system and does not explicitly specify the goal to be attained.

In this paper we show that the joint learning of the sensory and the control part produces a system that is robust with respect to unmodeled disturbances. This is a critical issue for stereo vision systems: for example an insult to the system might cause a displacement of one camera, which in turn modifies the extrinsic parameters (i.e., the relative offset of the two cameras) of the model of the system. We consider four different types of perturbations that we apply to one camera: blur, roll (in-plane rotation), tilt (vertical misalignment), and pan (horizontal misalignment). We show that the system can still learn vergence despite the perturbations. Moreover, when a perturbation is introduced, adapting the bases of the sparse coding models to the changed input statistics improves the performance, as compared to a case where only the policy of the RL agent is adapted and the bases are tuned to unperturbed images. The results underline the importance of adapting both the sensory encoding and the behavior of the system. The use of an intrinsic reward, coupled to an efficient coding of the sensory inputs, allows the model to continuously learn under a multitude of conditions. This self-calibrating property is highly desirable for robotic systems that have to operate in changing environments.

We use the head of the humanoid robot iCub as a test platform. The iCub robot stereo head represents a convenient platform to study active perception, because it replicates the main degrees

of freedom of the human head and eyes. We train the model using the iCub simulator and use it to quantitatively assess the performance of the system. We then show that the model also works well on the real robot. The paper is organized as follows: in section 2 we describe the model architecture, the perturbations used and the experimental setup. Section 3 contains the results of the robustness analysis and section 4 discusses the results.

2. MATERIALS AND METHODS

In this section, we first provide an overview of the architecture of the vergence control system; then we describe the set of distortions applied to the stereo images, which are used to assess the robustness of the method. Finally we describe the iCub robotic platform and the simulator used to run the experiments.

2.1. MODEL ARCHITECTURE

The vergence control model consists of three main stages (see **Figure 1**):

- Pre-processing: stereo patches are extracted from the input binocular images and normalized.
- Sensory encoding: two sparse coding models are used to encode the input images at different resolutions.
- Motor control: a reinforcement learning agent generates vergence commands to move the cameras of the robot according to the output of the sparse coding models.

A detailed description of our model architecture has been introduced in (Lonini et al., 2013c). We report here the main elements for the sake of completeness.

2.1.1. Pre-processing

Stereo images are acquired from the cameras of the iCub robot (320×240 pixels) and converted to gray-scale. The fixation point is defined to be at the center of each input image. A 128×128 pixel image is cut from the center of left and right images (**Figure 1**, red windows); the image is subsampled to 16×16 pixels using a Gaussian pyramid and patches of size 8×8 pixels are extracted; this set of patches (receptive fields) corresponds to patches of size 64×64 in the original image. The subsampling operation is performed to reduce the computational burden required to train the sparse coding model as well as to learn basis functions at a coarse resolution. To learn basis functions at a fine resolution, patches of size 8×8 pixels are extracted from 72×72 pixel foveal windows (**Figure 1**, blue windows), without performing any subsampling. From each foveal window, we extract a total of 81 patches of size 8×8 pixels, where patches at the coarse scale are shifted horizontally and/or vertically by multiples of 1 pixel. This ensures that the same number of patches is extracted at each scale. For each scale, each left (right) patch is transformed into a column vector x_k^L (x_k^R) and preprocessed to have zero mean and unit norm. Corresponding left and right patches are vertically concatenated to form a stereo-patch x_k , where the first 64 components of x_k correspond to the left patch and the last 64 correspond to the right patch. The subscript k indexes the patch within an image.

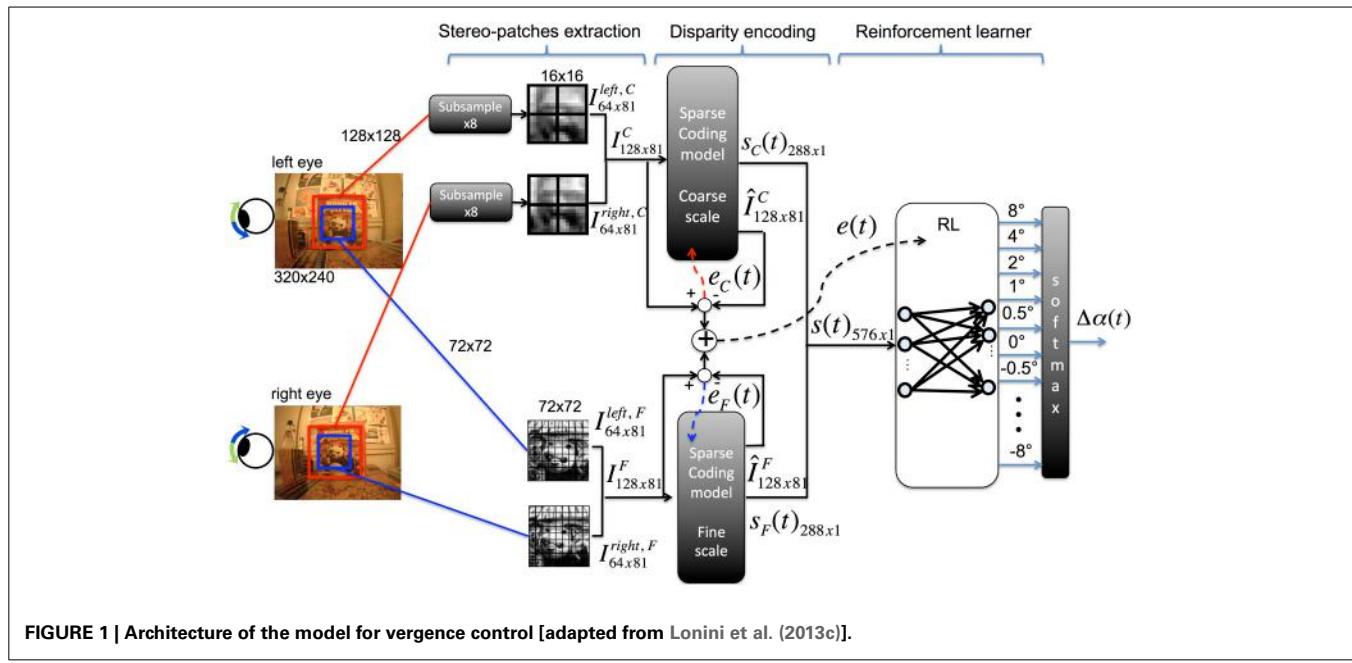


FIGURE 1 | Architecture of the model for vergence control [adapted from Lonini et al. (2013c)].

2.1.2. Sparse coding model

The input to a sparse coding model is a matrix of 81 patches within an input stereo image at a given scale (i.e., coarse or fine). A stereo patch is approximated through the sparse coding model by a linear combination of binocular (stereo) basis functions ϕ . Formally this approximation is expressed by:

$$\begin{bmatrix} \hat{x}_k^L \\ \hat{x}_k^R \end{bmatrix} = \sum_{i=1}^B a_i^{(k)} \begin{bmatrix} \phi_i^L \\ \phi_i^R \end{bmatrix}, \quad (1)$$

where $B = 288$ is the total number of basis functions available in the dictionary of each sparse coding model. In order to ensure sparseness of the representation we allow only 10 coefficients a_i to be non-zero. The sparse coding model is trained to represent the original image as accurately as possible given this sparseness constraint. The total squared reconstruction error over all the stereo-patches, normalized by the energy in the original image measures the loss of information due to the encoding. This is defined by:

$$e = \frac{\sum_{k=1}^P \|x_k - \hat{x}_k\|^2}{\sum_{k=1}^P \|x_k\|^2}, \quad (2)$$

where P is the total number of patches within an image.

Learning occurs online through a two-step procedure: for each patch, a set of coefficients a_i and basis functions ϕ are selected from the basis dictionary using matching pursuit (Mallat and Zhang, 1993), a greedy algorithm that finds a set of bases to represent the input patch. Then, the chosen bases are adapted through gradient descent on the reconstruction

error function (Olshausen et al., 1996). Given a foveal window $\hat{I}(t)$ at time t and scale j (i.e., coarse or fine), we compute the B -dimensional feature vector, $s_j(t)$, by averaging the squared weighting coefficients over the P patches taken from the window:

$$s_j(t) = \begin{bmatrix} \frac{1}{P} \sum_{k=1}^P \left(a_1^{(k)}(t) \right)^2 \\ \vdots \\ \frac{1}{P} \sum_{k=1}^P \left(a_B^{(k)}(t) \right)^2 \end{bmatrix}, \quad (3)$$

where $a_i^{(k)}$ denotes the coefficient¹ of basis i for patch k .

In biological terms, each entry of the state vector models the pooled responses of binocular simple cells (coefficients $a_i^{(k)}$ for a given i) over different locations of the visual field (different patches k). The receptive field of a binocular simple cell is represented here by a basis function ϕ_i , which is sensitive to a specific orientation, spatial frequency and disparity. The result of this pooling roughly corresponds to the operation performed by complex cells, which receive inputs from many simple cells at different locations and tuned to the same disparity.

2.1.3. Reinforcement Learning

The reinforcement learning agent receives as input the combined feature vector $s(t)$ from each scale and maps it to a vergence change $\Delta\alpha(t)$. The reward for the agent is the negative sum of the reconstruction errors of the two sparse coding models. The goal of the RL agent is to select actions to maximize the discounted

¹For the convenience of reading, we drop the index j indicating that the coefficients a_i depend on the scale.

cumulative future reward $R(t)$

$$R(t) = \sum_{k=0}^{\infty} -\gamma^{-k}[e_C(t+k) + e_F(t+k)], \quad (4)$$

where e_C and e_F are the reconstruction errors (2) for the coarse and fine scale sparse coding models, respectively².

The RL architecture we use is the natural actor-critic algorithm as described in Bhatnagar et al. (2009), with an additional regularization factor to keep the weights of the policy bounded. Two linear neural networks (NN) are used to implement the actor (policy) and the critic (value function). The critic network receives as an input the state $s(t)$ and produces as output the value $V(t)$ of the current state

$$V(t) = v^T(t)s(t), \quad (5)$$

where $v(t)$ are the weights of the network at time t and the superscript T denotes the transpose operator. The policy network maps states to actions and its output layer contains as many neurons as possible actions that the agent can generate. Each action is a relative change $\Delta\alpha(t)$ in the current vergence angle $\alpha(t)$. We chose a set A of 11 actions, uniformly spaced on a logarithmic scale as $A = \{-8^\circ, -4^\circ, -2^\circ, -1^\circ, -0.5^\circ, 0^\circ, 0.5^\circ, 1^\circ, \dots, 8^\circ\}$ to allow coarse and fine movements.

The activation z_a of each output neuron at time t is computed as

$$z_a(t) = \theta_a^T(t)s(t), \quad (6)$$

where $\theta_a(t)$ is the vector of weights from the state s to action a at time t .

The probability of choosing action a is computed according to a softmax operation on the activation of the output neurons that is:

$$\pi_a(s(t)) = \frac{\exp(\beta z_a(t))}{\sum_{j=1}^{11} \exp(\beta z_j(t))}, \quad (7)$$

where β is the inverse of the temperature parameter which controls the amount of exploration vs. exploitation. During training this parameter is set to 1.

2.2. IMAGE PERTURBATIONS

We consider four types of perturbations to assess the robustness as well as the adaptation properties of the model. These perturbations simulate either an unmodeled disturbance or the consequence of an event which causes a change in the extrinsic camera parameters (e.g., a collision). The perturbations are simulated by applying the following transformations to one of the cameras of the robot (we chose the right one):

- Blur: the original image is blurred by applying a rotationally symmetric Gaussian lowpass filter. Three different levels of blur

are chosen, corresponding to the following three different combinations of the standard deviation σ and kernel size S of the filter reported in **Table 1**.

- Rotations: We add a constant roll (5° , 15° or 25°), tilt (2° , 6° or 16°) or pan (2° or 4°) angle to the right camera. The roll simulates an in-plane rotation of the camera; the tilt and pan mainly produce, respectively, a vertical and horizontal offset of the right image with respect to the left image. In biological terms, the pan and tilt rotations have a loose analogy with the clinical condition named strabismus, where the gaze direction of one eye is constantly deviated with respect to that of the other eye. In a robotic system this perturbation might occur as a result of an insult to the system.

The effect of each perturbation is shown in **Figure 2B**. Details on how to simulate those rotations from the original images are provided in the Appendix. Importantly, since the RL agent can only change the vergence angle, tilt and roll perturbations can not be fully compensated. In contrast, the effect of a pan perturbation can be fully compensated by the model through the controlled degree of freedom. We assess how the model deals with each condition.

2.3. EXPERIMENTAL SETUP

The iCub robot is an open source humanoid robotic platform. The head of the robot (Beira et al., 2006) has a total of six degrees of freedom: three in the neck (pan, tilt, roll) and three in the eyes (independent pan for left and right eye, common tilt). In our setup, we keep the neck of the robot fixed and control anti-symmetrically the pan of the two eyes such as to only modify the vergence angle. In order to accurately quantify the performance of our method, we train the model using the iCub simulator, which provides a controlled environment for extensive testing.

Table 1 | Parameters of the different blur levels.

σ [px]	4	16	32
S [px]	8×8	32×32	64×64

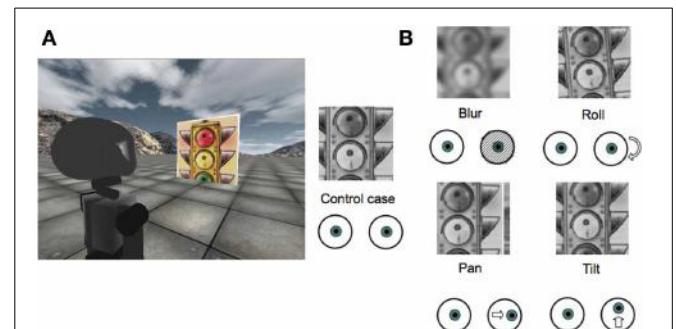


FIGURE 2 | (A) A screenshot from the iCub simulator showing the experimental setup; **(B)** Types of perturbations applied to right image from the robot camera (Blur of $\sigma = 4$ px; Roll of 5° ; Pan of 4° ; Tilt of 2°).

²Maximizing (4) corresponds to minimizing the total reconstruction error.

The stereo images acquired from the cameras of the simulated robot have a resolution of 320×240 pixels. The focal length is equivalent to 257 pixels which yields a horizontal field of view of $\sim 64^\circ$. Thus, a patch at the coarse and fine scale subtends a visual angle of, respectively, 14.2 and 1.8° .

We use a flat square object of side 1 m fronto-parallel to the robot at a varying distance ranging from 0.5 to 2 m (**Figure 2A**). During training the object distance is varied uniformly within that range every 10 iterations. This range of distances corresponds to vergence angles varying from 8 to 2° . We constrain the maximum vergence angle to be 20° . Similarly, the texture applied on the object is also changed by randomly drawing it from a set of 24 different images. Changing the texture provides the sparse coding model with sufficient statistics about the environmental stimuli to allow a diverse set of basis functions to develop. Training is performed online, where the sparse coding model as well as the RL are both updated at each iteration of the algorithm.

3. RESULTS

We first compare how performance changes when a distortion is present, with respect to the control model (i.e., a model trained without any distortion). Each model is trained for 100,000 iterations and performance is measured by the absolute mean vergence error (AME) during training. Since the largest action that the model can take in one step corresponds to a change of 8° in vergence, more than one step may be required to reach the target vergence value. For example, if the current vergence is 20° and the target vergence is 1° , the minimum number of steps required to reach the target vergence is 4 (one possible sequence of actions is $-8^\circ, -8^\circ, -2^\circ, -1^\circ$). In order to prevent a bias in the estimation of the performance, we only consider the error in the iteration preceding the stimulus change (i.e., the 9th iteration after presentation of a new stimulus). If the new stimulus is introduced at time t , the AME is

$$\text{AME}(t) = \frac{1}{N} \sum_{k=-N/2+1}^{N/2} |\alpha(t + 9 + 10k) - \alpha^*(t + 9 + 10k)| \quad (8)$$

where α^* is the target vergence angle for the stimulus and N is the size of the averaging window. In our experiments we use a value of $N = 500$ iterations. Since the averaging window is centered on the data point, to compute the AME when there are no previous or subsequent data points available (i.e., $t < N/2$ and $t > T - N/2$, with T being the total number of training iterations) we replicate the data point³.

Figure 3 shows the AME during training for four different perturbations, averaged over five different simulation runs. The level of the perturbation that we use corresponds to the images of **Figure 2**. As we can see from the decrease of the vergence error, the model can learn to verge under all types of perturbations considered. As a comparison, a random policy for selecting actions would lead to a vergence error of 7.5° . The performance of the system and its final accuracy depend both on the type of perturbation and, as we will show below, on its level. The AME for the

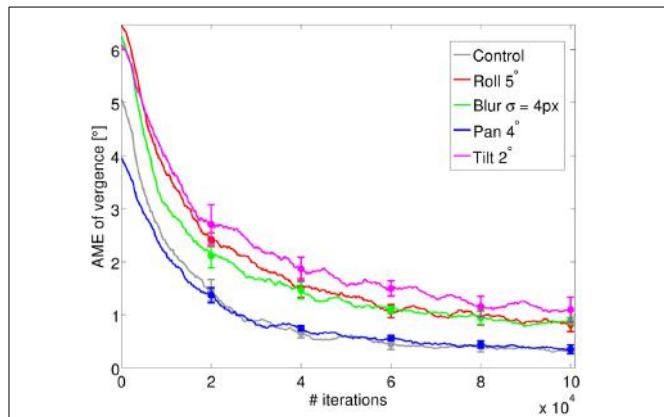


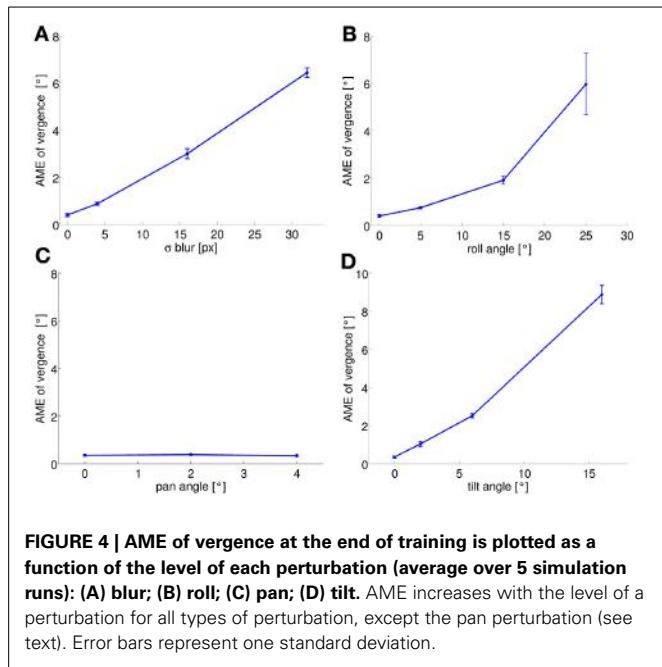
FIGURE 3 | The AME of vergence during training for the four types of perturbations introduced (blur, roll, pan, and tilt), averaged over 5 simulations. The control represents the case where no perturbation is applied (gray line). The AME for a random policy is $\sim 7.5^\circ$ (not shown). Error bars represent one standard deviation.

control settles at $\sim 0.2^\circ$ at the end of training. For the pan rotation (horizontal misalignment) the model displays a similar performance. This is because the system can still find a position of zero disparity and maximum redundancy by acting on the vergence angle. The vergence position in this case will correspond to the fixation on a point that is horizontally shifted by 2° with respect to the fixation point of the control case. This vergence position can be reached without any change in the system since our RL agent outputs relative vergence angles. For the other three perturbations the final accuracy is lower compared to the control case. In the case of blur, this is due to the loss of high frequency information. On the other hand, the tilt and roll rotations induce a change in the redundancy of information between left and right image at the vergence position, which affects the performance. However, the final error is $\sim 1^\circ$, which shows good learning of the vergence control.

As previously mentioned, the performance of the system at the end of learning also depends on the level of the perturbation we introduce. To quantify this performance we run the training phase with different levels of perturbation and observe the final AME of the vergence. **Figure 4** shows the results for each type of perturbation, averaged over 5 simulations as before. Again, the AME for a random policy is $\sim 7.5^\circ$. As expected, the performances are not affected in the pan rotation case. For the other conditions, the performance degrades as the level of each perturbation increases. In the case of blurred images, the learned policy performs better than a random policy up to values of $\sigma = 16$ pixels (AME $\sim 3^\circ$). For roll angles up to 15° , the AME is $\sim 2^\circ$, indicating that the model can learn vergence, despite the significant rotation between left and right images. For a roll angle of 25° the AME reaches on average 6° . The AME for the tilt perturbation reaches a value of $\sim 2.5^\circ$ for a tilt angle of 6° , which corresponds to a vertical offset of 18 pixels in the image. When the tilt angle is 16° (vertical offset of 74 pixels), performance degrades drastically and the AME increases to $\sim 8.5^\circ$.

In order to assess whether the model can generalize well on new data after training, a test is conducted using a new set of

³For example we assume $\alpha(t') = \alpha(0)$ if $t' < 0$.

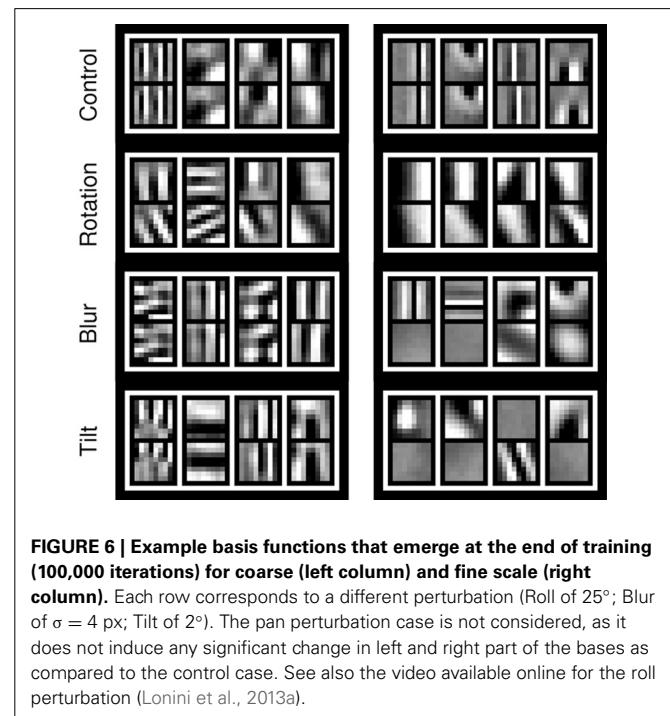
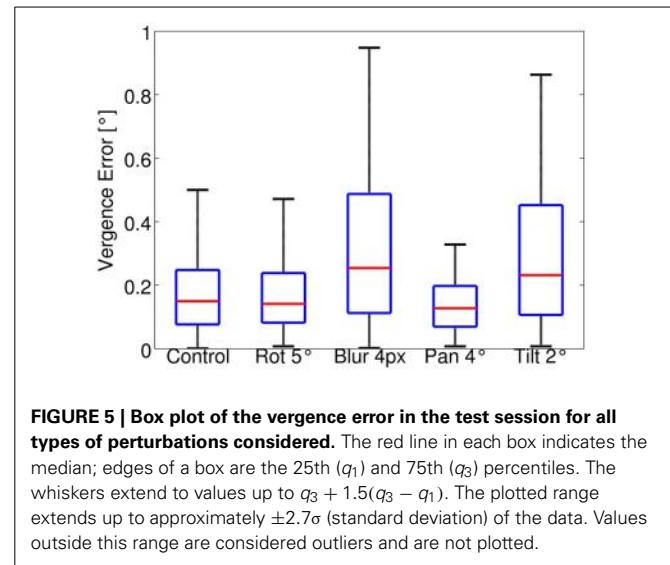


five textures and evaluating the greedy policy. In that case (7) is replaced by

$$\pi_a(s(t)) = \begin{cases} 1 & \text{if } a = \text{argmax}_a\{z_a(t)\} \\ 0 & \text{otherwise} \end{cases}. \quad (9)$$

The stimulus depth is randomly changed between 2 and 0.5 m every 10 iterations and a texture is drawn from the test set every 200 iterations. The test runs for a total of 1000 iterations. The same random sequence is used for all the perturbations. **Figure 5** shows a box plot of the vergence error during the test for each case. The median vergence error is used to remove the effect of outliers during the test. For the control case the median of the vergence error is 0.15° . The effect of a roll rotation of 5° is also fully compensated, while the blur ($\sigma = 4$ px) and the tilt rotation induce a slightly larger median vergence error, which is $\sim 0.25^\circ$. Overall the model performs well in the test sessions for all cases considered. In general, the errors are smaller than that measured at the end of training because the greedy policy is used for testing.

Figure 6 shows example basis functions from the learned dictionaries for each perturbation condition and for each scale. Basis functions are tuned to different orientations and spatial frequencies. Left and right part (vertically concatenated) for bases tuned to zero disparity are identical, while bases tuned to non-zero disparities show a horizontal shift between the left and right part (**Figure 6-Control**). Each perturbation induces a specific change in the bases that reflects the type of perturbation. The blur condition produces mostly monocular bases at the fine scale, indicated by the fact that the right part is plain. The roll perturbation induces a rotation of the right part with respect to the left, while the tilt rotation produces some bases with a vertical shift between left and right parts, representing vertical disparity.



To assess how adaptation of the bases affects learning of the policy when a perturbation is introduced, we consider the following scenario: we first train a model without any perturbation (control case) for 100,000 iterations. Then, a perturbation is introduced and the model is further trained under either of the following two conditions: first, the bases of the sparse coding models are updated and second, the bases remain fixed as they were before the perturbation. These situations may be roughly analogous to the biological case of an insult to the system occurring either before or after the end of the critical period (Hubel and Wiesel, 1970). In terms of robotics, this could correspond to a perturbation induced by a shock received by the robot, after

the system has been trained in an unperturbed scenario. **Figure 7** shows the AME during training for three different perturbations (blur, roll and tilt perturbation, first row) as well as the reconstruction error of the sparse coding model for the fine scale (bottom row), under the two conditions. We observe that the AME decreases more for the case where the bases are allowed to change vs. the case when the system uses the same bases learned in the no-perturbation condition (**Figure 7**, red vs. blue line). Importantly, the policy weights are allowed to change in both cases. Thus, the RL can adapt to the perturbation, even when the same set of basis functions is used. As expected, when the bases are allowed to change, the reconstruction error decreases. This is because the adapted bases can represent the perturbed images better than the original set of bases, trained on unperturbed images; moreover the policy that emerges leads to lower vergence errors, which translates into lower reconstruction errors. Notably, the reconstruction error for the blur case drops in both conditions (adapting and non-adapting bases) because blurring one of the images makes it easier to encode. Also, the AME for the roll perturbation at the onset of the perturbation ($\sim 3^\circ$) is lower than the AME obtained at the end of training for the same type of perturbation (cfr. **Figures 4B, 7**). The reason is that the bases trained in absence of the perturbation can still be used to detect disparity, when the perturbation is introduced. A video showing the development of the basis functions, before and after the roll perturbation is introduced, is available online (Lonini et al., 2013a). It can be seen that during exposure to the perturbation, the right part of several basis functions rotates, relatively to the left part.

Finally, we test the model trained in the simulator on the real robot to assess the performance when different perturbations are applied. Three sources of uncertainties affect the reliability of the measure of the vergence error on the iCub: 1) the backlash in the DC motors ($\leq 1^\circ$) that prevents us from accurately measuring the actual vergence angle from the encoder readings; 2) the error in

the measure of the distance of the stimulus from the robot; 3) the estimates of the extrinsic camera parameters as well as lens distortions. **Figure 8** shows the left and right image anaglyph from the robot cameras before and after vergence is achieved, for all types of perturbations (blur of $\sigma = 4$ px; roll of 5° ; pan of 4° ; tilt of 2°). The model is able to achieve correct vergence under all the perturbations considered. Of notice, the camera parameters of the real iCub differ from that of the simulator. A video of the robot performing the vergence in each condition is available online (Lonini et al., 2013b).

4. DISCUSSION

Despite an increasing interest in intrinsic motivations there is still no universally accepted definition. One standpoint is that extrinsic motivations are driven by variables outside of the controller (e.g., battery level, state of the sensors), whereas intrinsic motivations are related to variables within the brain (or controller) of the agent. Thus, intrinsic motivations are driven by epistemic goals, i.e., goals directed to improve the knowledge of the agent, rather than producing a direct change in the world (Baldassarre, 2011). (Zhao et al., 2012) and (Lonini et al., 2013c) have recently proposed a form of intrinsically motivated learning for efficient coding in active perception. They generalize classic notions of efficient coding to movements of the sense organs that facilitate efficient encoding of the sensory data. To this end, a sensory coding model is coupled with a reinforcement learner for controlling the sense organs. The reinforcement learner is rewarded for movements that make the sensory input easier to encode. This approach is closely related to a recent formulation of intrinsic motivations as aiming to maximize compression progress (Schmidhuber, 2009) to create a more compact (and thus interesting) representation of the data. Our system also favors compression progress because achieving a smaller reconstruction error after a vergence command, while using the same amount of neural resources (number of active basis functions),

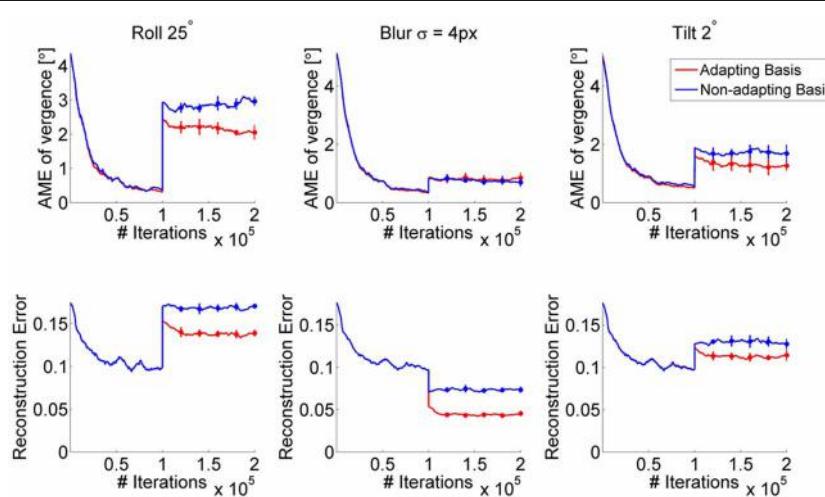


FIGURE 7 | AME (first row) and reconstruction error of the fine scale sparse coding models (bottom row) when the bases of the sparse coding models are adapted (red) vs. fixed (blue).

The perturbation occurs after 100,000 iterations. Curves show the average over 5 simulations. Error bars represent one standard deviation.



FIGURE 8 | Test on the real iCub. Anaglyph of left and right image, before (left) and after (right) vergence is achieved for each perturbation (from top to bottom: Blur of $\sigma = 4$ px; Roll of 5° ; Pan of 4° ; Tilt of 2°). See also online video (Lonini et al., 2013b).

implies that the data are encoded more efficiently. Zhao et al. (2012) and (Lonini et al., 2013c) have shown that in the context of binocular vision, this leads to a fully autonomous learning of disparity representations and accurate vergence control. The system discovers that it is *useful* to properly verge its eyes, because this enables it to encode the sensory data more efficiently.

In this paper we build on this previous work and provide an analysis of the robustness of the approach to various perturbations. We believe that the robustness and self-calibrating properties of a robotic system are a matter of great importance when building autonomous robots capable of adapting to changing environments. We first show that learning occurs under all the perturbations considered and the model performance degrades gracefully with the size of the perturbation. We then compare the condition where the bases (filters) are allowed to adapt when a perturbation is present with the case where they are left unchanged from training on normal images. Adaptation of the bases leads to a more efficient encoding of the input images, which

in turns leads the RL to adapt the policy, in a completely unsupervised fashion. Thus, a changed condition in the system, such as a rotation or misalignment of a camera, is automatically handled by our model. A complete compensation of the pan perturbation is obtained as the model controls the vergence angle. Similarly, a full compensation for the tilt and roll perturbation could be achieved if the RL agent was allowed to independently control the tilt and roll angle for each eye.

Previous work addressing the issue of vergence in active stereo vision systems has often relied on computer vision techniques to infer disparity from the stereo pair, and then controlling the stereo cameras through a feedback loop. These methods often require the knowledge of the intrinsic (e.g., focal length and optical centers of the cameras) and the extrinsic (relative position of the two cameras) parameters of the cameras. Examples include cepstral or zero-disparity filters (Olson and Coombs, 1991), correlation-based methods (Capurro et al., 1997) and feature matching (Hansen and Sommer, 1996). Reinforcement learning has been used to learn vergence, by using as reward the disparity estimated through feature matching (Piater et al., 1999) or by a population of disparity-tuned neurons (Franz and Triesch, 2007; Wang and Shi, 2010). The main limitation of these approaches is that the disparity filters are not learned from the data. Importantly, to our knowledge, there is no work that is directly addressing the robustness of a vergence control method to image distortions.

Our model provides a way to autonomously adapt both the sensory representation as well as the control of the behavior by the simultaneous learning of the two systems. The proposed method can be extended to other domains, such as the learning of smooth-pursuit behavior, which is currently under development. Future work should address whether this new framework for efficient coding in active perception can be further extended to other sensory modalities and what insights into the biology of active perception it provides.

ACKNOWLEDGMENTS

The authors wish to thank Pramod Chandrashekhariah for helpful discussions.

FUNDING

This work has partly received funding from the European Community's Seventh Framework Programme FP7/2007-2013, Challenge 2—Cognitive Systems, Interaction, Robotics" under grant agreement No FP7-ICT-IP-231722, project IM-CLeVeR Intrinsically Motivated Cumulative Learning Versatile Robots, from the Hong Kong RGC and the German DAAD through the Germany/Hong Kong Joint Research Scheme (project number G_HK25/10) and from the BMBF Project "Bernstein Fokus: Neurotechnologie Frankfurt, FKZ 01GQ0840."

REFERENCES

- Atneave, F. (1954). Some informational aspects of visual perception. *Psychol. Rev.* 61, 183. doi: 10.1037/h0054663
- Baldassarre, G. (2011). "What are intrinsic motivations? a biological perspective," in *ICDL 2011, IEEE international conference on Development and Learning and Epigenetic Robotics*, Vol. 2, (Frankfurt am Main: IEEE) 1–8.
- Barlow, H. B. (1961). "Possible principles underlying the transformation of sensory messages," in *Sensory Communication*, eds W. Rosenblith (Cambridge: MIT Press), 217–234.

- Beira, R., Lopes, M., Praga, M., Santos-Victor, J., Bernardino, A., Metta, G., Becchi, F., and Saltarén, R. (2006). "Design of the robot-cub (iCub) head," in *ICRA 2006, IEEE International Conference on Robotics and Automation* (Orlando, FL: IEEE), 94–100.
- Bhatnagar, S., Sutton, R. S., Ghavamzadeh, M., and Lee, M. (2009). Natural actor-critic algorithms. *Automatica* 45, 2471–2482. doi: 10.1016/j.automatica.2009.07.008
- Capurro, C., Panerai, F., and Sandini, G. (1997). Dynamic vergence using log-polar images. *Int. J. Comp. Vis.* 24, 79–94. doi: 10.1023/A:1007974208880
- Faugeras, O. D. (1993). *Three Dimensional Computer Vision: A Geometric Viewpoint*. Cambridge, MA: MIT Press.
- Field, D. J. (1994). What is the goal of sensory coding? *Neural Comput.* 6, 559–601. doi: 10.1162/neco.1994.6.4.559
- Franz, A. and Triesch, J. (2007). "Emergence of disparity tuning during the development of vergence eye movements," in *ICDL 2007. IEEE 6th International Conference on Development and Learning* (London: IEEE), 31–36.
- Hansen, M. and Sommer, G. (1996). "Active depth estimation with gaze and vergence control using gabor filters," in *Proceedings of the 13th International Conference on Pattern Recognition*, Vol. 1, (Vienna: IEEE), 287–291. doi: 10.1109/ICPR.1996.546035
- Hubel, D. H. and Wiesel, T. N. (1970). The period of susceptibility to the physiological effects of unilateral eye closure in kittens. *J. Physiol.* 206, 419.
- Lennie, P. (2003). The cost of cortical computation. *Curr. Biol.* 13, 493–497. doi: 10.1016/S0960-9822(03)00135-0
- Lonini, L., Forestier, S., Teuliére, C., Zhao, Y., Shi, B. E., and Triesch, J. (2013a). *Basis Functions Adaptation during Exposure to a Perturbation*. Available online at: <http://youtu.be/dOqNJngh84U>
- Lonini, L., Forestier, S., Teuliére, C., Zhao, Y., Shi, B. E., and Triesch, J. (2013b). *Robust Active Binocular Vision Through Intrinsically Motivated Learning on iCub*. Available online at: <http://youtu.be/hcbxzgrYdlo>
- Lonini, L., Zhao, Y., Chandrashekhariah, P., Shi, B. E., and Triesch, J. (2013c). "Autonomous learning of active multi-scale binocular vision," in *ICDL 2013, IEEE International Conference on Development and Learning and Epigenetic Robotics*, (Osaka: IEEE).
- Mallat, S. G. and Zhang, Z. (1993). Matching pursuits with time-frequency dictionaries. *IEEE Trans. Signal Process.* 41, 3397–3415. doi: 10.1109/78.258082
- Olshausen, B. A. et al. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* 381, 607–609. doi: 10.1038/381607a0
- Olson, T. J. and Coombs, D. J. (1991). Real-time vergence control for binocular robots. *Int. J. Comp. Vis.* 7, 67–89. doi: 10.1007/BF00130490
- Perez-Orive, J., Mazor, O., Turner, G. C., Cassenaer, S., Wilson, R. I., and Laurent, G. (2002). Oscillations and sparsening of odor representations in the mushroom body. *Science* 297, 359–365. doi: 10.1126/science.1070502
- Piater, J. H., Grupen, R. A., and Ramamritham, K. (1999). "Learning real-time stereo vergence control," in *Proceedings of the 1999 IEEE International Symposium on Intelligent Control/Intelligent Systems and Semiotics*, (Cambridge, MA: IEEE), 272–277.
- Rothkopf, C. A., Weisswange, T. H., and Triesch, J. (2009). "Learning independent causes in natural images explains the spacevariant oblique effect," in *ICDL 2009. IEEE 8th International Conference on Development and Learning*, (IEEE) 1–6.
- Schmidhuber, J. (2009). "Driven by compression progress: a simple principle explains essential aspects of subjective beauty, novelty, surprise, interestingness, attention, curiosity, creativity, art, science, music, jokes," in *Anticipatory Behavior in Adaptive Learning Systems Lecture Notes in Computer Science*, Vol. 5499, eds G. Pezzulo, M. V. Butz, O. Sigaud, and G. Baldassarre (Berlin, Heidelberg: Springer), 48–76.
- Smith, E. C. and Lewicki, M. S. (2006). Efficient auditory coding. *Nature* 439, 978–982. doi: 10.1038/nature04485
- Wang, Y. and Shi, B. E. (2010). Autonomous development of vergence control driven by disparity energy neuron populations. *Neural Comput.* 22, 730–751. doi: 10.1162/neco.2009.01-09-950
- Weliky, M., Fiser, J., Hunt, R. H., and Wagner, D. N. (2003). Coding of natural scenes in primary visual cortex. *Neuron* 37, 703–718. doi: 10.1016/S0896-6273(03)00022-9
- Willshaw, D. J., Buneman, O. P., and Longuet-Higgins, H. C. (1969). Non-holographic associative memory. *Nature* 222, 960–962. doi: 10.1038/222960a0
- Zhao, Y., Rothkopf, C. A., Triesch, J., and Shi, B. E. (2012). "A unified model of the joint development of disparity selectivity and vergence control," in *ICDL 2012, IEEE International Conference on Development and Learning and Epigenetic Robotics* (San Diego, CA: IEEE), 1–6. doi: 10.1109/DevLrn.2012.6400876

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 11 July 2013; accepted: 10 October 2013; published online: 07 November 2013.

*Citation: Lonini L, Forestier S, Teuliére C, Zhao Y, Shi BE and Triesch J (2013) Robust active binocular vision through intrinsically motivated learning. *Front. Neurorobot.* 7:20. doi: 10.3389/fnbot.2013.00020*

*This article was submitted to the journal *Frontiers in Neurorobotics*.*

Copyright © 2013 Lonini, Forestier, Teuliére, Zhao, Shi and Triesch. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

APPENDIX

We describe here how the perturbations are generated. A rotation of a camera induces a projective transformation of the image. To simulate our pan, tilt and roll perturbations we thus compute this transformation, also called *homography*. Formally, the homographic image transformations H , is computed by:

$$H = KRK^{-1}, \quad (10)$$

where K is a 3×3 matrix, containing the camera intrinsic parameters (focal length and image center coordinates) and R is a 3×3 matrix containing the three angles of rotation of the camera (pan, tilt, in-plane rotation⁴) (Faugeras, 1993). The forms of K and R are the following:

$$K = \begin{bmatrix} f & 0 & c_x \\ 0 & f & c_y \\ 0 & 0 & 1 \end{bmatrix}, \quad (11)$$

$$R = \begin{bmatrix} c(r_y)c(r_z) & -c(r_y)s(r_z) & s(r_y) \\ c(r_x)s(r_z) + s(r_x) & c(r_x)c(r_z) - s(r_x)s(r_y) & -s(r_x)c(r_y) \\ \sin(r_y)c(r_z) & s(r_x)s(r_y)s(r_z) & c(r_x)c(r_y) \\ c(r_x)s(r_y)c(r_z) + s(r_x)\sin(r_z) & c(r_x)s(r_y)s(r_z) + c(r_z)s(r_x) & +c(r_z)s(r_x) \end{bmatrix}, \quad (12)$$

where c_x and c_y are the image center coordinates, f is the focal in pixel values and $c()$ and $s()$ denote, respectively, the cosine and sine operation. r_x , r_y , and r_z indicate the tilt, pan and roll angle, respectively. We thus simulate those rotations by applying the homographic transformation to the acquired images. For each pixel of coordinates $[uv]$ of the original image, the corresponding position after the perturbation would be $[u'v'1]^\top = H[uv1]^\top$. Cubic interpolation is used to compute the pixel values at integer pixel coordinates. Remember that the system has no knowledge about the perturbation nor the camera parameters, and the steps described here are only used to simulate a perturbation in one of the cameras. Also, as the result of a rotation is a relative misalignment of the two cameras, it is actually irrelevant whether the rotation is applied to one or both cameras to demonstrate the ability of the system to correct for a perturbation.

⁴We indicate the in-plane rotation degree of freedom also as *roll*.



The role of intrinsic motivations in attention allocation and shifting

Dario Di Nocera^{1*}, Alberto Finzi^{2*}, Silvia Rossi^{2*} and Mariacarla Staffa^{2*}

¹ Dipartimento di Matematica e Applicazioni, Università degli Studi di Napoli "Federico II," Napoli, Italy

² Dipartimento di Ingegneria Elettrica e Tecnologie dell'Informazione, Università degli Studi di Napoli "Federico II," Napoli, Italy

Edited by:

Gianluca Baldassarre, Italian National Research Council, Italy

Reviewed by:

Alexander Förster, University of Applied Sciences and Arts of Southern Switzerland, Switzerland
Nazmul Haque Siddique, University of Ulster, UK

***Correspondence:**

Dario Di Nocera, Dipartimento di Matematica e Applicazioni, Università degli Studi di Napoli "Federico II," Comp. Univ., Monte S. Angelo, via Cinthia 26, I-80126 Napoli, Italy
e-mail: dario.di.nocera@gmail.com;

Alberto Finzi, Silvia Rossi and
Mariacarla Staffa, Dipartimento di Ingegneria Elettrica e Tecnologie dell'Informazione, Università degli Studi di Napoli "Federico II," via Claudio 21, 80125 Napoli, Italy
e-mail: alberto.finzi@unina.it; silvia.rossi@unina.it; mariacarla.staffa@unina.it

INTRODUCTION

Attention and *intrinsic motivations* play a crucial role in cognitive control (Posner et al., 1980) and are of great interest in cognitive robotics. Indeed, attentional mechanisms and motivational drives are strictly involved in the process of guiding and orchestrating multiple concurrent behaviors. Attentional mechanisms, beyond their role in perception orientation, are also considered as key mechanisms in action selection and coordination (Posner et al., 1980; Norman and Shallice, 1986). The capability of selecting and filtering the information is associated with the process of focusing cognitive and executive resources toward the stimuli that are relevant for the environmental and behavioral context. On the other hand, another key factor affecting action selection is represented by the so called intrinsic motivations such as the curiosity (Baldassarre, 2011), which can indirectly affect action selection because of its influence on attentional shifting. For instance, the curiosity drive can attract the attentional focus toward novel stimuli and, consequently, can elicit the execution of actions which are not directly related to the current behavior or goal. Albeit there is not a clear consensus on how intrinsic motivations differ from the extrinsic ones (Baldassarre, 2011), their role in pushing human/animal beings to spontaneously explore their environment (Baldassarre and Mirolli, 2013) and to execute this activity only for their inherent satisfaction (Ryan and Deci, 2000), rather than for satisfying some basic needs such as hunger or thirst (White, 1959; Berlyne, 1960), is widely accepted.

The concepts of attention and intrinsic motivations are of great interest within adaptive robotic systems, and can be exploited in order to guide, activate, and coordinate multiple concurrent behaviors. Attention allocation strategies represent key capabilities of human beings, which are strictly connected with action selection and execution mechanisms, while intrinsic motivations directly affect the allocation of attentional resources. In this paper we propose a model of Reinforcement Learning (RL), where both these capabilities are involved. RL is deployed to learn how to allocate attentional resources in a behavior-based robotic system, while action selection is obtained as a side effect of the resulting motivated attentional behaviors. Moreover, the influence of intrinsic motivations in attention orientation is obtained by introducing rewards associated with curiosity drives. In this way, the learning process is affected not only by goal-specific rewards, but also by intrinsic motivations.

Keywords: attention shifting, curiosity, intrinsic motivations, reinforcement learning, action selection

In this work, we focus on the intrinsic motivation provided by the curiosity, which is considered as the main drive for humans to explore novel situations and to learn complex behaviors from experience (Berlyne, 1954; Litman, 2005). Recent studies have also shown that both attention and curiosity are strictly related to the dopaminergic system responsible for action driving. It is widely accepted, indeed, that dopamine affects both the reward excitement, fundamental in the learning process, and the demand of more attention by novel stimuli (Nieoullon, 2002; Redgrave and Gurney, 2006; Jepma et al., 2012). Unpredicted events can generate intrinsic reinforcement signals, which support the acquisition of novel actions. In particular, it has been shown that the dopamine release is triggered not only in response to unexpected environmental changes and goal-directed action-outcome learning (Heidbreder and Groenewegen, 2003; Dalley et al., 2004), but also in response to the detection of novel events (Lisman and Grace, 2005).

The typical approach adopted for modeling the dopamine-like rewarding system (Montague et al., 1996) and for coping with the problem of treating intrinsic motivations (Barto et al., 2004; Mirolli and Baldassarre, 2013) is represented by the well known Reinforcement Learning (RL) process. Recent works have been proposed to incorporate models for novelty (Marsland et al., 2000) and curiosity (Schmidhuber, 1991) within Motivated RL algorithms (Barto et al., 2004) providing accounts for behavior adaptation, action selection learning, mental development,

and learning of hierarchical collections of skills depending on the robot experience (Kaplan and Oudeyer, 2003; Barto et al., 2004; Oudeyer and Kaplan, 2007; Schembri et al., 2007; Singh et al., 2010; Baranes and Oudeyer, 2013). Typically, within these approaches, RL is used to directly model and generate the action selection strategies. In contrast, we propose a system where RL is deployed to learn attentional allocation and shifting strategies, while action selection emerges from the regulation of attentional monitoring mechanisms (Di Nocera et al., 2012), which can be affected by the intrinsic motivation of curiosity. Our curiosity model is inspired by the interest/deprivation model proposed by Litman (2005), which captures both optimal-arousal and curiosity-driven approaches of curiosity modeling. Following this approach, attentional shifting mechanisms can be generated taking into account not only extrinsic motivations, like mission goals and primary needs satiation, but also intrinsic motivations, like the need of acquiring knowledge (Litman's deprivation model) and the attraction toward novel stimuli and opportunity of learning (Litman's interest-based model). In this context, we aim to investigate whether our account of curiosity and attentional regulation learning is feasible and effective for the generation of attentional allocation and shifting strategies, whose side effect is an adaptive emergent behavior for the robot. We are also interested in the impact of our model of curiosity on the learning process. Specifically, we want to assess whether the proposed intrinsically motivated system affects the progress in learning.

We detail the approach by describing our intrinsically motivated RL model and analyzing its performance in a simulated survival domain. In this scenario, the robot is engaged in survival tasks such as finding food or water, while avoiding dangerous situations. The goal is to learn attentional allocation and shifting policies that allow the robot to survive in the particular environment. The system evaluation is based on a comparison between the performance of the attentional policies, which are learned with the curiosity model, with respect to the ones generated without taking into account the curiosity drive. The collected results show that our intrinsically motivated learning approach is feasible and effective. Indeed, the curiosity-driven learning system allows us to find satisfactory attentional allocation and shifting policies showing a faster convergence of the learning process, safer policies of the selected action, and a higher wellness state of the robotic system in terms of energy gained during the exploration of the environment. In particular, in the curious setting the robot behavior seems more flexible because endowed with an additional capacity of adaptation. Indeed, different attentional allocation (or shifting) policies, and consequently, various action selection policies, can be defined depending on the current level of curiosity.

MATERIAL AND METHODS

ATTENTIONAL SHIFTING SYSTEM

In this work, we refer to the attentional framework introduced by Burattini et al. (2010). Here, the attentional system is modeled as a reactive behavior-based system (Brooks, 1986; Arkin, 1998), endowed with internal attentional mechanisms capable of distributing and shifting the attention among different concurrent behaviors depending on the current saliency of tasks

and stimuli. These attentional mechanisms allow the robotic system to supervise multiple concurrent behaviors and to efficiently manage limited resources. In contrast with typical works on visual attention (Itti and Koch, 2001), the Burattini et al. (2010) approach is not concerned with the orientation of the attention in the space (i.e., the field of view), but it is about the executive attention (Posner et al., 1980) and the temporal distribution of the attentional resources needed to monitor and control multiple processes. This model of attention is inspired by Pashler and Johnston (1998), where the attentional load due to the accomplishment of a particular task is defined as the quantity of attentional time units devoted to that particular task, and by Senders (1964), where attentional allocation and shifting mechanisms are related to the sampling rate needed to monitor multiple parallel processes. In particular, Burattini et al. (2010) propose a frequency-based model of attention allocation, where the increment of the attention due to salient stimuli is associated with an increment of sensors sampling rate and of the behavior activations. Specifically, starting from a behavior-based architecture, each behavior is endowed with an internal clock regulating its activation frequency and sensory sampling rate: the higher the sampling rate, the higher the resolution at which the behavior is monitored and controlled. The internal clock can increase or decrease the attentional state of each behavior with respect to salient internal/external stimuli by means of suitable attentional monitoring functions. In this context, the internal stimuli are modeled as internal needs, such as, for example, thirst or hunger, while the external stimuli are associated with salient events or discontinuities perceived in the external environment.

An explicative example of this behavior-based attentional system at work is presented in **Figure 1**. The plot shows how the sampling rate of a behavior (for example a *give* task) changes (see **Figure 1A**) depending on different stimuli (for example, the human hand speed, in **Figure 1B**, and the distance between the human hand and the robot end effector, in **Figure 1C**). It is possible to observe that if non-salient stimuli are presented to the behavior, the attentional process monitors the environment in a relaxed manner, instead, if something salient happens, the clock frequency of the behavior is enhanced and more attention is consequently paid toward the stimulus. This general model permits to monitor and control different internal and external processes, shifting, from time to time, the allocation of computational and operational resources. Notice that, this adaptive frequency implicitly provides a mechanism for behaviors prioritization. Indeed, high-frequency behaviors are associated with activities with a high relevance and priority in the current operational context.

Formalization of the model

Following the approach of Burattini and Rossi (2008), we consider a Behavior-based architecture (Brooks, 1986; Arkin, 1998), where each behavior is endowed with an attentional mechanism represented by an internal adaptive clock.

A schema theory representation (Arbib, 1998) of an attentional behavior is illustrated in **Figure 2**. This is characterized by a Perceptual Schema (PS), which elaborates sensor data, a Motor Schema (MS), producing the pattern of motor actions, and an

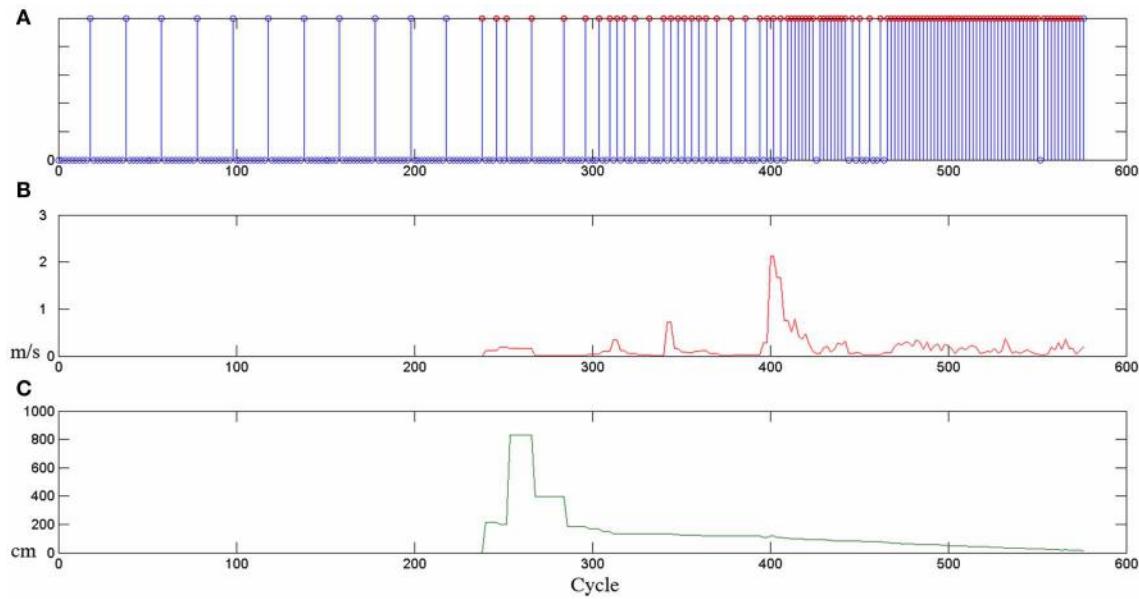


FIGURE 1 | Example of the attentional allocation strategies presented in Sidobre et al. (2012). (A) The sampling rate associated with the behavior of giving an object changes depending on internal or external stimuli; **(B)** the human hand speed and **(C)** the distance between the human hand and the robot end-effector.

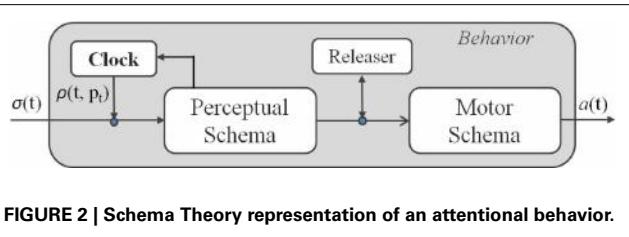


FIGURE 2 | Schema Theory representation of an attentional behavior.

attentional control mechanism, called Adaptive Innate Releasing Mechanism (AIRM), based on a combination of a clock and a releasing mechanism works as a trigger for the MS activation (e.g., the view of a predator releases the escape behavior), while the clock regulates the sensors sampling rate and, consequently, the activation rate of the behaviors. The clock activation rate changes following an attentional monitoring strategy, which can adaptively increase or decrease the clock frequency, according to salient internal and external stimuli. More formally, the attentional mechanism is characterized by:

- An activation period p_t^b ranging in an interval $[p_{min}^b, p_{max}^b]$, where b is the behavior identifier.
- A *monitoring function* $f(\sigma^b(t), p_{t-1}^b) : \mathbb{R}^n \rightarrow \mathbb{R}$ that adjusts the current clock period p_t^b , according to the internal state of the behavior and to the environmental changes.
- A trigger function $\rho(t, p_t^b)$, assuming a 0/1 value, which enables/disables the data flow $\sigma^b(t)$ from sensors to PS at each p_t^b time unit.
- Finally, a normalization function $\phi(f(\sigma^b(t), p_{t-1}^b)) : \mathbb{R} \rightarrow \mathbb{N}$ that maps the values returned by f into the allowed range $[p_{min}^b, p_{max}^b]$.

The clock period at time t is regulated as follows:

$$p_t^b = \rho(t, p_{t-1}^b) \cdot \phi(f(\sigma^b(t), p_{t-1}^b)) + (1 - \rho(t, p_{t-1}^b)) \cdot p_{t-1}^b \quad (1)$$

That is, if the behavior is disabled, the clock period remains unchanged, i.e., p_{t-1}^b . Otherwise, when the trigger function is 1, the behavior is activated and the clock period changes according to the $\phi(f)$. In order to learn attentional monitoring strategies, various methods such as Differential Evolution (Burattini et al., 2010) and RL techniques (Di Nocera et al., 2012) have been deployed, respectively for off-line and on-line tuning of the parameters regulating the attentional monitoring functions. In the following sections, we will present an intrinsically motivated RL (IMRL) approach to the attentional allocation problem in our frequency-based model of attention.

INTRINSIC MOTIVATIONS: CURIOSITY MODEL

Curiosity is an appetitive state involving the recognition, pursuit, and intense desire to investigate novel information and experiences that demand one's attention. In literature, we find two main theoretical accounts of curiosity: the *optimal arousal model* and *curiosity-drive theory*. The curiosity-drive model assumes that the main drive of curiosity is the reduction of uncertainty: novel and ambiguous stimuli cause a need for coherence restore that reduces the uncertainty. This reduction is considered as rewarding. This model is supported by studies showing that unusual situations are associated with approaching behaviors and attentional states (e.g., see the Loewenstein, 1994 knowledge gap/approach gradient). However, the curiosity-driven model cannot explain why biological organisms initiate exploratory behaviors without any stimuli. These situations are instead well explained by the optimal-arousal model (e.g., see the Spielberger and Starr, 1994 model). Following

this model, the biological systems are associated with an homeostatic regulation of their arousal level: when the arousal level is under-stimulated, the organism is motivated to increase the arousal and to look for novel situations; in contrast, when the organisms is over-stimulated additional stimuli are evaluated as negative and associated with an avoidance behavior. While in the curiosity-drive model the reward is associated with uncertainty reduction, in the optimal arousal model, the induction of curiosity is directly rewarding. Also this model is not completely satisfactory, indeed, in this case an optimal arousal state should be maintained and the reward is directly associated with a feeling of interest, hence the gain of new knowledge could reduce this feeling and could be considered counter-productive (see Litman, 2005 for a discussion). A combination of these two approaches is proposed by Litman (2005) with the *interest/deprivation* model of curiosity. Here, both the satiation and the activation of curiosity can be rewarding: the interest-based curiosity is driven by novel stimuli and opportunity of learning, whereas the deprivation-based curiosity is driven by the uncertainty and the lack of knowledge. In Litman (2005) the interest/deprivation model of curiosity is then related to the neuroscience of the wanting and liking systems, which are hypothesized to underlie motivation and affective experience for a wide class of appetites (Berridge, 2003). In the Litman model, wanting is associated with deprivation and need of knowledge, while liking is associated with the expected pleasure due to learning and knowledge acquisition. In Table 1 we show the Litman's classification. Here, in the case of high level of wanting and liking, curiosity is due to a need of knowledge and it is sustained by an interest; if wanting is low, but liking is high, information seeking is motivated by pure interest; in contrast, if wanting is high and liking is low, the need of knowledge is not associated with the anticipation of a pleasure. Finally, when wanting and liking are low, also the curiosity drive is inhibited. In this paper, we exploit a model of curiosity that is inspired by the Litman (2005) interpretation of wanting and liking.

REINFORCEMENT LEARNING FOR ATTENTIONAL SHIFTING

Following the approach by Di Nocera et al. (2012), in this paper we exploit a RL algorithm to learn the attention allocation strategies introduced in section 2.1. In Di Nocera et al. (2012), a Q-learning algorithm is used to tune and adapt the frequencies of sensors sampling, while action selection is obtained as a side effect of this attentional regulation. In the following, we first recall the Q-learning algorithm and then we detail its application to the attentional shifting problem.

Table 1 | Litman's classification of curiosity states with respect to high and low levels of liking and wanting (Litman, 2005).

Liking	Wanting	
	Low	High
Low	LL: Ambivalent disinterest	LH: Need for uncertainty clarification
High	HL: Curiosity as a feeling of "interest"	HH: Curiosity as a feeling of "deprivation"

General description of the Q-learning algorithm

Q-learning (QL) (Watkins and Dayan, 1992) is a learning algorithm for a Markov Decision Process (MDP). A MDP is defined by a tuple (S, A, R, P_a) where S is the set of states, A is the set of actions, R is the reward function $R : S \times A \rightarrow \mathbb{R}$, with $R(s, a)$ the immediate reward in $s \in S$ after the execution of $a \in A$; P_a is the transition function $P_a : S \times A \times S \rightarrow [0, 1] \in \mathbb{R}$, with $P_a(s, a, s')$ probability of $s' \in S$ after the execution of $a \in A$ in $s \in S$. A solution of a MDP is a policy $\pi : S \rightarrow A$ that maps states into actions. The *value function* $V^\pi(s)$ is the cumulative expected reward from $s \in S$ following π . The *q-value* $Q(s, a)$ is the expected discounted sum of future payoffs obtained by executing the action a from the state s and following an optimal policy π^* , i.e., $Q(s, a) = \{R_{t+1} + \gamma V^*(s_{t+1}) \mid s_t = s, a_t = a\}$, with V^* associated to π^* .

In QL techniques, the Q-values are estimated through the agent experience after being initialized to arbitrary numbers. For each execution of an action a_t leading from the state s_t to the state s_{t+1} , the agent receives a reward r_{t+1} , and the Q-value is updated as follows:

$$Q(s_t, a_t) \leftarrow (1 - \alpha_t) \cdot Q(s_t, a_t) + \alpha_t (R_{t+1} + \gamma \cdot \max_{a_{t+1} \in A} Q(s_{t+1}, a_{t+1})) \quad (2)$$

where γ is the discount factor (which determines the importance of future rewards) and α is the learning rate.

Different exploration policies can be introduced to select the action to be executed trying to balance exploration and exploitation. Analogously to Di Nocera et al. (2012), in this paper we consider a *Softmax* method that selects the action to be executed through a Boltzmann distribution (Sutton and Barto, 1998):

$$P_a(a \mid s, Q) = \frac{\exp^{\frac{Q(s,a)}{\tau}}}{\sum_{b \in A(s)} \exp^{\frac{Q(s,b)}{\tau}}} \quad (3)$$

Here, the temperature τ controls the exploration strategy: the higher the temperature, the closer the strategy is to a random policy (exploration); the lower the temperature, the closer the strategy is to $Q(s, a)$ maximization (exploitation). Under suitable conditions (see, for example, Watkins and Dayan, 1992), this algorithm converges to the correct Q-values with probability 1 assuming that every action is executed in every state infinitely many times and α is decayed appropriately.

Q-learning for attentional regulation

In our setting, the QL algorithm is to be exploited to generate the attention allocation strategy. For this purpose, we introduce a suitable space state S^b for each attentional behavior, while the action space A^b represents a set of possible regulations of the clocks. Specifically, the action space spans a discretized set of possible allowed periods $P^b = \{p_1^b, \dots, p_k^b\}$ for each behavior b (i.e., A^b coincides with P^b). Since the current state $s^b \in S^b$ should track both the attentional state (clock period) and the perceptive state, this can be represented by a pair $s^b = (p^b, \sigma^b)$, where $p^b \in P^b$ is the current clock period and $\sigma^b \in X^b$ is for the current perceptive status. In particular, we consider the perceptive state of each

behavior as a discretization of the behavior perceptive domain using n equidimensional intervals $X^b = \{\sigma_1^b, \dots, \sigma_n^b\}$. Therefore, the attentional allocation policy $\pi^b : S^b \rightarrow A^b$ represents a mapping between the current state s^b and the next attentional period p^b that should be learned by means of the QL algorithm. That is, given a reward function R for each behavior, the algorithm is to find the optimal attention allocation policy π^b , i.e., for each state $s^b \in S^b$, the activation period $p^b \in P^b$ that maximizes the expected reward of that behavior.

The resulting Q-table for a generic attentional behavior in Di Nocera et al. (2012) can be described by the **Table 2**.

This approach to adaptive attentional allocation and action selection has been tested in a robotic setting (Di Nocera et al., 2012). Starting from this model we will design our model of *Intrinsically Motivated Reinforcement Learning*.

MOTIVATED RL FOR ATTENTIONAL SHIFTING

In this section, we extend the RL approach to attention allocation presented above introducing the effects of the intrinsic motivation of curiosity. In particular, we rely on a curiosity model which is inspired by the interest/deprivation model proposed by Litman (2005) and adapted to the behavior-based setting we consider in this work. More specifically, analogously to Litman, we associated the liking mechanism to a direct reward related to novelty, however, our interpretation of the wanting system is slightly different. Indeed, in the place of the cognitive deprivation model introduced by Litman, which cannot be easily accounted within the simple behaviors we are concerned with, we relate the wanting mechanism to the need to explore and act. This is represented by a value that we called the residual energy: the higher the available energy, the higher is the need to “consume” this surplus in an exploratory (hence, curious) behavior. More details will be provided below and in section 4 where we present some concrete instances of the reward functions used to capture this model of curiosity.

ACTION SPACE

Analogously to Di Nocera et al. (2012), in our model, for each behavior b we introduce an *Action Space* A^b representing the set of possible periods $P^b = \{p_1^b, \dots, p_k^b\}$ for that behavior. That is, an action a^b is a possible assignment of a clock period p^b which regulates the sampling rate and the activation frequency of the

associated behavior. As explained above, the idea is that the system does not learn directly the action to execute, instead, it learns the attentional policies (i.e., clock regulations with respect to its perceptual and attentional state). In this context, the action selection is an indirect consequence of the attentional behaviors. In the curiosity-driven setting, different attentional shifting strategies will be learned depending on the level of curiosity of the agent.

STATE SPACE

In order to represent the curiosity state into the state space, we reformulate the *State Space* S^b of Di Nocera et al. (2012) introducing a new parameter representing the degree of curiosity of the agent. In the extended framework, the state s^b is determined by a triple (c^b, p^b, σ^b) , where, c^b represents the level of curiosity of the system, p^b is for the current clock period, and σ^b is the current perceptive state of a behavior b . In particular, for each behavior, the attentional monitoring period p^b ranges in a predefined set of possible values P^b . Analogously, the perceptive state σ^b is suitably discretized in intervals representing sub-ranges of the input signal X^b . Finally, the curiosity degree c^b ranges in an interval of the four values $[LL, LH, HL, HH]$ representing four relations between *wanting* and *liking* values (low-low, low-high, high-low, high-high) which are inspired by the curiosity model definition introduced in Litman (2005) (see section 2.2). Therefore, the attentional allocation policy $\pi^b : S^b \rightarrow A^b$ represents a mapping between the current state s^b and the next action a^b corresponding to the suitable period for the attentional monitoring p^b , that is learned by means of the QL algorithm.

REWARD FUNCTION

Given the Q-Learning Actions and States Spaces, we can introduce the *Reward function* as a combination of *extrinsic* component, an *intrinsic* component and a dynamic weight between these two. While the extrinsic reward depends on the direct effect of the actions with respect to the behavior utility, in our curiosity model, the second reward is directly related to the pleasure of the novelty, hence to the level of liking. Instead, the wanting level is used to dynamically balance the relation between extrinsic and liking reward: the higher the need of information seeking, the higher the liking associated with the encountered novelty. As stated before, differently from Litman (2005), our assumption is that the level of wanting depends on a sort of (global) energy state of the agent (see section 4 for additional details in the case study). The idea is that the robotic agent can explore new situations, guided by curiosity, only when the system is in a wellness state. Instead, when the system is under a certain wellness threshold, the attention is focused on priority needs (e.g., to eat and drink) rather than on secondary ones (information seeking and exploration of new states). We formalize the overall reward function as follows:

$$R^b = (1 - w) \cdot R_e^b + (w) \cdot R_l^b \quad (4)$$

where R_e^b is the reward computed considering the observed state, and R_l^b represents the reward evaluated considering the satisfaction of an observation with respect to a particular curiosity state (i.e., the reward is related to something that the agent likes just

Table 2 | Q-values for a generic behavior, where S^b represents the state space.

S^b		A^b			
		p_1	p_2	...	p_k
σ_1	p_1	$Q_{11,1}$	$Q_{11,2}$...	$Q_{11,k}$

	p_k	$Q_{1k,1}$	$Q_{1k,2}$...	$Q_{1k,k}$
\dots
	p_1	$Q_{n1,1}$	$Q_{n1,2}$...	$Q_{n1,k}$

σ_n	p_k	$Q_{nk,1}$	$Q_{nk,2}$...	$Q_{nk,k}$

because it is novel). The value of R_l^b is thus computed as level of *liking*. The w value represents the level of *wanting*, an internal unmotivated need to explore something (the drive toward a specific location/object depends on the liking mechanism).

This relation between liking and extrinsic rewards implies that, when the situation is critical (i.e., low energy) the R_l reward value will be neglected with respect to the R_e extrinsic reward value, while R_l will be emphasized as much as the agent will be in a wellness state. The possible correspondences between the R_l , R_e rewards and the *wanting*, *liking* values are illustrated in the **Table 3**. Notice that this matrix of wanting and liking relations is different from the one by Litman (2005), because of the different interpretation of the wanting system. For example, here low wanting and liking levels are associated with the prevalence of extrinsic rewards, while in Litman (2005) they are directly associated to a boredom state.

CASE STUDY

In order to test our approach we introduce a *Survival Domain*, where a robot must survive for a predefined amount of time within an environment (see **Figure 3**) avoiding obstacles (objects, walls, etc.) and recharging energy by eating and drinking.

We consider simulated environments of 16 m^2 . Obstacles, water, and food locations are cubes of size $0.5 \text{ m} \times 0.5 \text{ m} \times 0.5 \text{ m}$, respectively of black, blue, and green color (see **Figure 3**). An experiment ends in a positive way if the robot is able to survive till the end of the test (*max_time*), while it fails in the following three cases: (1) the robot collides with an obstacle or its distance from an obstacle is under a certain *safety* distance threshold; (2) the value representing the robot *thirst* goes under the minimum value; (3) the value representing the *hunger* goes under the minimum value. We tested our approach using a simulated *Pioneer3-DX* mobile robot (using the Player/Stage tool Gerkey et al., 2003), endowed with a blob camera and 16 sonar sensors.

Internal needs functions

We assume that the robot is endowed with internal drives. In our case study, we consider two internal needs: hunger and thirst. These are modeled by the following functions.

We introduce a *Hunger function*, to compute the need for food:

$$\begin{aligned} \text{Hunger}(t) = & \text{Hunger}(t - 1) + k \cdot (\text{nb_act}) \\ & - (e_f \cdot \text{food_consumed}) \end{aligned} \quad (5)$$

Here, the hunger increases the need for food at each machine cycle by a k value, for each active behavior (nb_act), and decreases it

Table 3 | Wanting and liking relations and the associations between liking and extrinsic rewards.

Liking	Wanting	
	Low	High
Low	$LL: R_e >> R_l$	$LH: R_e < R_l$
High	$HL: R_e > R_l$	$HH: R_e << R_l$

when a quantity of food is ingested (*food_consumed*), depending on the energy power of the food (e_f).

An analogous *Thirst function* is used to compute the need for water:

$$\begin{aligned} \text{Thirst}(t) = & \text{Thirst}(t - 1) + k \cdot (\text{nb_act}) \\ & - (e_w \cdot \text{water_consumed}) \end{aligned} \quad (6)$$

ATTENTIONAL BEHAVIOR-BASED ARCHITECTURE

We introduce a Behavior-Based Attentional Architecture (see **Figure 4**) where, the attentional control is obtained from the interaction of a set of three parallel attentional behaviors AVOID, EAT and DRINK.

For each behavior, the process of changing the rate of sensory readings is interpreted as an increase or decrease of selective attention toward internal or external saliences. These sources of salience are generally behavior- and task-dependent; these can depend on either internal states, such as hunger, thirst, etc., or external stimuli, such as obstacles, unexpected variations of the environment, attractiveness of a particular object, etc. The overall attentional behavior should emerge from the interrelation of the attentional mechanisms associated with the different primitive behaviors and learned by means of the motivated RL learning technique.

In **Figure 4** we illustrate the attentional control system designed for the survival domain. It combines three behaviors: AVOID, EAT, and DRINK, each endowed with its releaser and

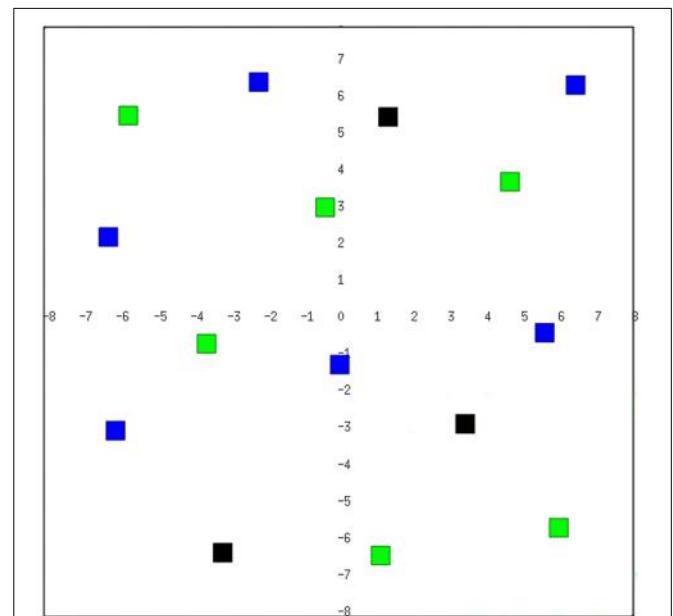


FIGURE 3 | The testing environment is simulated through the Player/Stage tool for robotics development (Gerkey et al., 2003). We adopt a simulated *Pioneer3-DX* mobile robot endowed with a blob camera and 16 sonar sensors. The black, blue, and green colored cubes (of size $0.5 \text{ m} \times 0.5 \text{ m} \times 0.5 \text{ m}$) within the environment represent respectively obstacles, water, and food.

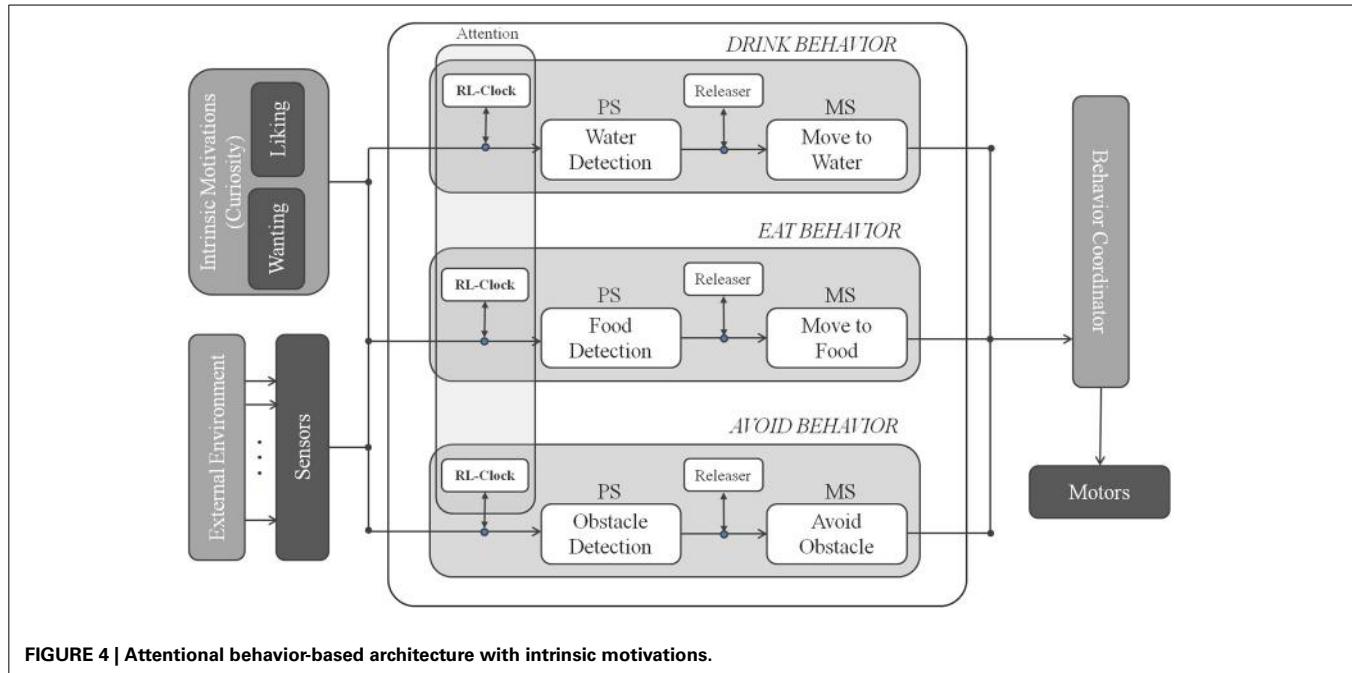


FIGURE 4 | Attentional behavior-based architecture with intrinsic motivations.

adaptive clock. The output of the robotic system is the combination of the outputs (if they are available).

AVOID

Manages obstacle avoidance. Its input signal σ_t^{avoid} is the minimum distance of the 8 frontal sonar sensors; its motor schema controls the robot linear and angular velocity ($v(t)$, $\theta(t)$) generating a movement away from the obstacle. The obstacle avoidance is obtained as follows: $v(t)$ is proportional to the obstacle proximity, i.e., $v(t) = v_{max} \cdot \frac{\sigma_t^{avoid}}{\sigma_{max}^{avoid}}$, where v_{max} and σ_{max}^{avoid} , are respectively the maximum velocity and the maximum sonar range; $\theta(t)$ is obtained as weighted sum of the angular velocities generated by the active sonars, i.e., $\theta(t) = \sum_{i \in A(t)} \theta_{max} \cdot \theta_i$, where $A(t)$ is the set of active sonars detecting an obstacle at time t , θ_{max} is the maximal rotation, θ_i is a suitable weight depending on the sonar position (high values for frontal sonars and low for lateral ones).

EAT

Monitors an internal function $Hunger(t)$ representing the need of food. At each execution cycle the Hunger function changes as described in the previous section. Therefore, EAT is active when $\sigma_t^{eat} = Hunger(t)$ goes above a suitable threshold σ_{max}^{eat} . When enabled, if a green blob (representing the food source) is detected by the camera, the motor schema generates a movement toward it, otherwise it starts looking around for the green, generating a random direction.

DRINK

Monitors a function $Thirst(t)$ that represents the need of water and considers the height (pixels in the field of view) of a detected blue object in the environment as an indirect measure of the distance from the object. The motor schema is enabled whenever the $\sigma_t^{drink} = Thirst(t)$ is greater than a suitable threshold σ_{max}^{drink} . When enabled, if a blue blob is detected by the camera, the

motor schema generates a movement toward it, otherwise it starts looking around as for the EAT behavior.

For each behavior, the clock regulation depends on the monitoring function that should be learned at run-time.

MOTIVATED ATTENTIONAL FRAMEWORK

Action and state spaces

In order to cast the RL problem in our case study, we have to define A^b 's and S^b 's. In our attentional allocation problem, for each behavior, the action space A^b is represented by a set of possible periods $\{p_1^b, \dots, p_k^b\}$ for the adaptive clock of each behavior b . In the case study, for each behavior (AVOID, EAT and DRINK) we assume 1 machine cycle as the minimum clock period and the following set of possible periods: $p^a, p^e, p^d = \{1, 4, 8, 12\}$. As for the state space S^b , we recall that each state is a triple (σ^b, p_i^b, c^b) composed of a value in the perceptual domain, a period, and a curiosity value. The perceptive state of each behavior is obtained as a discretization in six equidimensional intervals of the perceptive domain $[\sigma_{min}^b, \sigma_{max}^b]$. The perceptive domain for AVOID spans the interval $[0, \sigma_{max}^{avoid}]$, where σ_{max}^{avoid} is maximum sonar range for the behavior; the domain of DRINK is $[0, \sigma_{max}^{drink}]$, where σ_{max}^{drink} represents the maximum value for the Thirst function; the EAT domain is in $[0, \sigma_{max}^{eat}]$, where σ_{max}^{eat} is the maximum state of hunger the robotic system can assume. The curiosity value ranges in the conceptual interval $[LL, LH, HL, HH]$, where the combination of the wanting and liking parameters is considered.

Rewards

We assume the reward always positive except for a strong *penalty* if the system cannot survive. For the other cases the reward is computed as follows. For each behavior, the extrinsic reward has two additive components. The first evaluates the impact of frequent activations of a specific behavior. The higher is the frequency, the smaller is the obtained reward. This component is

equal to zero if $p_t^b = p_{min}^b$. The second component depends on the specific behavior.

In particular, concerning AVOID, each activation is rewarded directly proportional to the distance from the obstacle.

$$R_e^{avoid}(t) = \begin{cases} \frac{1}{2} \cdot \left[\left(\frac{p_t^{avoid} - p_{min}^{avoid}}{p_{max}^{avoid} - p_{min}^{avoid}} \right) + \left(\frac{\sigma_t^{avoid} - \sigma_{min}^{avoid}}{\sigma_{max}^{avoid} - \sigma_{min}^{avoid}} \right) \right], & \text{if !crash} \\ \text{penalty}, & \text{otherwise} \end{cases} \quad (7)$$

As for EAT behavior, for each activation the reward is inversely proportional to the current hunger value. That is, a system that is more hungry takes a smaller reward.

$$R_e^{eat}(t) = \begin{cases} \frac{1}{2} \cdot \left[\left(\frac{p_t^{eat} - p_{min}^{eat}}{p_{max}^{eat} - p_{min}^{eat}} \right) + \left(1 - \frac{\sigma_t^{eat}}{\sigma_{max}^{eat}} \right) \right], & \text{if !crash} \\ \text{penalty}, & \text{otherwise} \end{cases} \quad (8)$$

Analogously, each activation of DRINK is rewarded in inverse proportion to the current value of thirst:

$$R_e^{drink}(t) = \begin{cases} \frac{1}{2} \cdot \left[\left(\frac{p_t^{drink} - p_{min}^{drink}}{p_{max}^{drink} - p_{min}^{drink}} \right) + \left(1 - \frac{\sigma_t^{drink}}{\sigma_{max}^{drink}} \right) \right], & \text{if !crash} \\ \text{penalty}, & \text{otherwise} \end{cases} \quad (9)$$

Following the description of section 3, we subdivide curiosity into two components dealing respectively with the feeling of *wanting* and *liking*. We associate the first one to the concept of *residual energy* for the robot body, while the second one to the level of *novelty* in the exploration of the learning states.

In particular, we assume that the *Energy* of the system is defined as follows:

$$E(t) = E(t-1) - e_u - e_{nb} \cdot (nb_act) + e_f \cdot (food_consumed) + e_w \cdot (water_consumed) \quad (10)$$

where the current value of the energy $E(t)$ is computed starting from the previous level of energy $E(t-1)$, decremented of one unit of energy e_u , which represents the energy consumed at each machine cycle. Then, we also consider the energy spent to activate each behavior e_{nb} , where nb_act is the number of currently active behaviors. On the other hand, we assume increments of the energy in correspondence of consummatory behaviors such as EAT or DRINK, where the added quantity e_f (e_w) of energy depends on the consumed food or water (this is added when boolean conditions related to *food_consumed* consumed and *water_consumed* consumed becomes true).

According to the model of curiosity considered in this paper, we model the level of the *wanting* component of the curiosity as the residue of the Energy value (see Figure 5) ranging within the interval [0,1].

$$w = \begin{cases} \frac{E(t) - E_{well}}{E_{max} - E_{well}}, & \text{if } E(t) \geq E_{well} \\ 0, & \text{otherwise} \end{cases} \quad (11)$$

That is, the robot can show a curious behavior only when the situation is not critic (i.e., only when the global energy exceeds

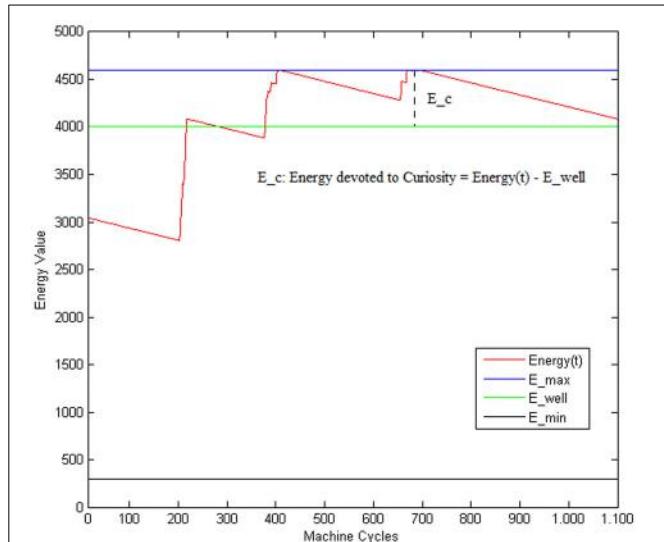


FIGURE 5 | $E(t)$ is the current Energy level; E_{min} : is the minimum amount of Energy permitting the system to work; E_{well} : is the level of Energy corresponding to a wellness state of the system.

the E_{well} threshold, indicating a sort of *wellness* state of the system). E_{well} is supposed to be associated with a state of the system where the regulation of the different behaviors activation periods is well balanced and leads to a suitable scheduling of the actions (reach food and water when necessary while avoiding obstacles). We can interpret this residual value E_c as the Energy that the system can spend on activities which are not associated with primary needs. In this way, the higher the E_c , the more the curiosity can drive the system to explore new states, the less the attention is posed on the primary behaviors (such as EAT, DRINK or AVOID). According to Equations (11, 4), c^b ranges only within an interval of three values [$LL - HL$, LH , HH]. LL and HL (i.e., both with low *wanting*) are considered as equivalent and correspond to the case of w equal to 0 (e.g., no curiosity).

The second component of the curiosity is the *liking*, which we associate with the pleasure due to novel situations. In particular, the curiosity in our system is interpreted as the exploration within the learning states space. We can assume that the novelty of a state is computed as follows:

$$R_l^b = 1 - \frac{NV(\sigma_t^b)}{NV_{tot}} \quad (12)$$

where, NV is for number of visits and $\frac{NV(\sigma_t^b)}{NV_{tot}}$ represents the number of times the percept σ_t^b has been observed during the previous NV_{tot} observations. We, thus, maintain a sort of temporal window of value NV_{tot} . In this way, on the one hand, we capture the novelty of the observation; on the other hand, we simulate a sort of lapsing mechanism where the novelty of a state is reduced when it is frequently visited within the time window. The model of the temporal window can be compared to the Itti's model of *surprise* (Baldi and Itti, 2010), by interpreting the temporal window as a rough approximation of a statistic on the perceptual history.

That is, if the system is not observing a percept for NV_{tot} times, the stimulus becomes likable again. While, if the system observes that particular perceptual state σ_t^b many times (i.e., NV_{tot}), the stimulus associated becomes boring. The R_l values range in the interval [0,1], so values greater than 0.5 indicates states with high *liking*.

The combination of *wanting* and *liking* drives model the curiosity which will affect the learning system explorative attitude.

Parameters and settings

In Table 4 we summarize the parameters and the settings used for our experiments.

Here, the perceptual domain (Perceptions) and the possible periods (PeriodsActions) are analogous to the ones presented in Di Nocera et al. (2012). Indeed, this partition for the perceptive domain and the periods have been selected to obtain a satisfactory setting for the non-curious system. As for curiosity, it is associated with the residual Energy with respect to a threshold set as the 2/3 of the maximum energy E_{max} . The maximum energy value E_{max} is set with respect to the *max_cycles* that estimates the maximum clock cycles associated with an episode (180 s to accomplish the survival task). This regulation is a compromise between scarce energy (that would keep the system in the non-curious state) and abundant energy (that would keep the system in the curious state). The minimal energy E_{min} is set to 300. Here, for each behavior activation we have an energy consumption of 1 *UoE* while the recharge is 150 *UoE* for food and water. Concerning the liking, we employed a temporal windows of 10 observations to assess the novelty of a perceptive data. As far as the learning parameters are concerned, we set $\alpha = 0.8$ for the learning rate, $\gamma = 0.9$ for the discount factor and $\tau = 1$ for the temperature. These regulations have been defined after a preliminary phase of experimental testing in the non-curious setting (analogous to the one presented in Di Nocera et al., 2012).

RESULTS

In this section, we present the experimental results of a robot that must survive for a predefined amount of time within an environment (see Figure 3) avoiding obstacles (objects, walls, etc.) and meeting its energy needs by eating and drinking. We

discuss the approach by considering the performance of the intrinsically motivated RL in learning attentional allocation policies in this survival domain. In particular, our aim is to evaluate the effects of the curiosity on the RL process by comparing the behavior of the curious system (from now on called CR = *CuriousRobot*) with respect the one of system that is not endowed with the curiosity drive (from now on called NCR = *Non-CuriousRobot*). Namely, the difference between the CR and NCR models is that the latter does not consider the rewards due to the curiosity. Notice that the parameter regulation process described in the previous section was carried out in order to obtain the best regulation for the non-curious system. Since these settings are shared by the curious and non-curious system, we can assess the added value of the intrinsically motivated framework in the testing scenario.

In order to evaluate how the curiosity affects the learning process, we first compare the survival time percentage of the CR with respect to the NCR. In Figure 6 we plot the survival time percentage averaged every 50 episodes.

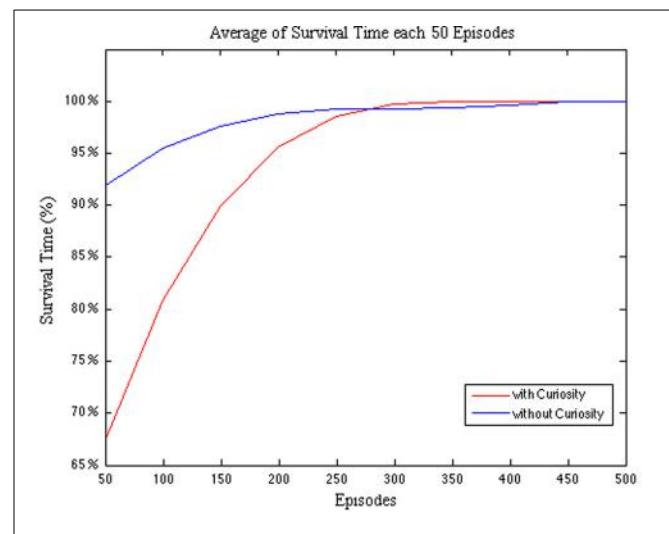


FIGURE 6 | Comparison between CR and NCR systems with respect to the survival time percentage per Episode. The survival time is averaged every 50 episodes.

Table 4 | Table of the parameters experimental setting (*UoE*, Unit of Energy; *UoR*, Units of Reward; *mc*, machine cycles; *m*, meters; *s*, seconds; *obs*, observations).

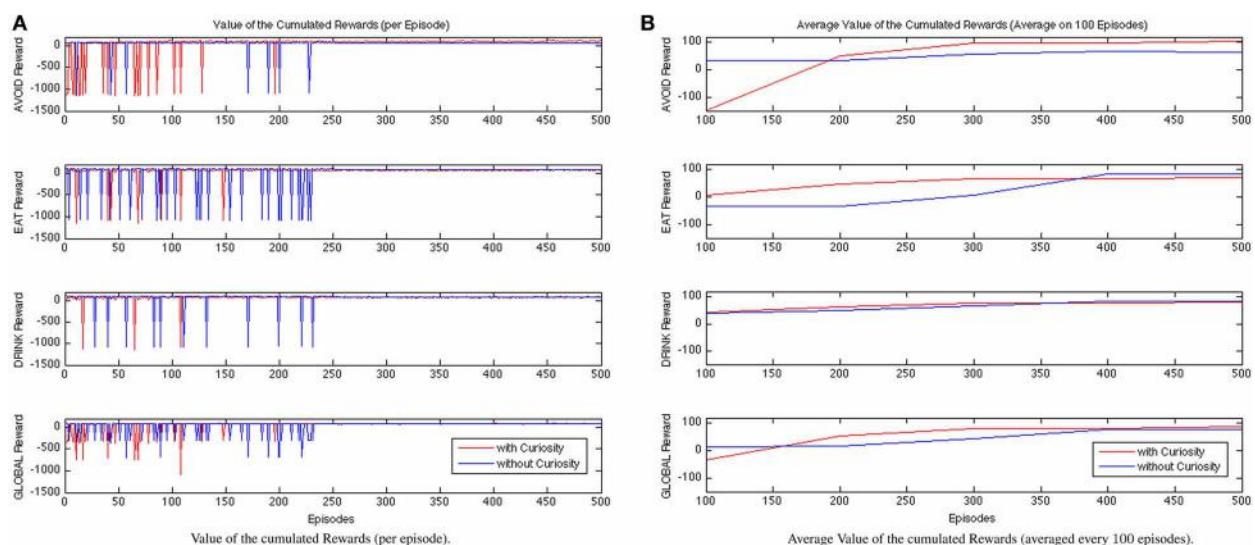
Experimental settings					
Perceptions		Curiosity		Episode	
σ_{max}^{avoid}	1.0 m	E_{max}	6000 <i>UoE</i> (4 * <i>max_cycles</i>)	<i>max_time</i>	180 s
σ_{min}^{avoid}	0.4 m	E_{min}	300 <i>UoE</i>	<i>max_cycles</i>	1500 <i>mc</i>
σ_{eat}^{max}	1500 <i>UoE</i>	E_{well}	(2/3)* E_{max}	<i>penalty</i>	-1500 <i>UoR</i>
σ_{eat}^{min}	300 <i>UoE</i>	NV_{tot}	10 obs	Power of Food/Water	
σ_{drink}^{max}	1500 <i>UoE</i>	α	0.8	e_f	150 <i>UoE</i>
σ_{drink}^{min}	300 <i>UoE</i>	γ	0.9	e_w	150 <i>UoE</i>
PeriodsActions		τ	1	e_u	1 <i>UoE</i>
p^a, p^e, p^d	{1 mc, 4 mc, 8 mc, 12 mc}			e_{nb}	1 <i>UoE</i>

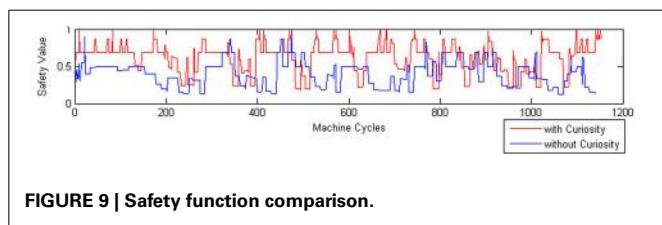
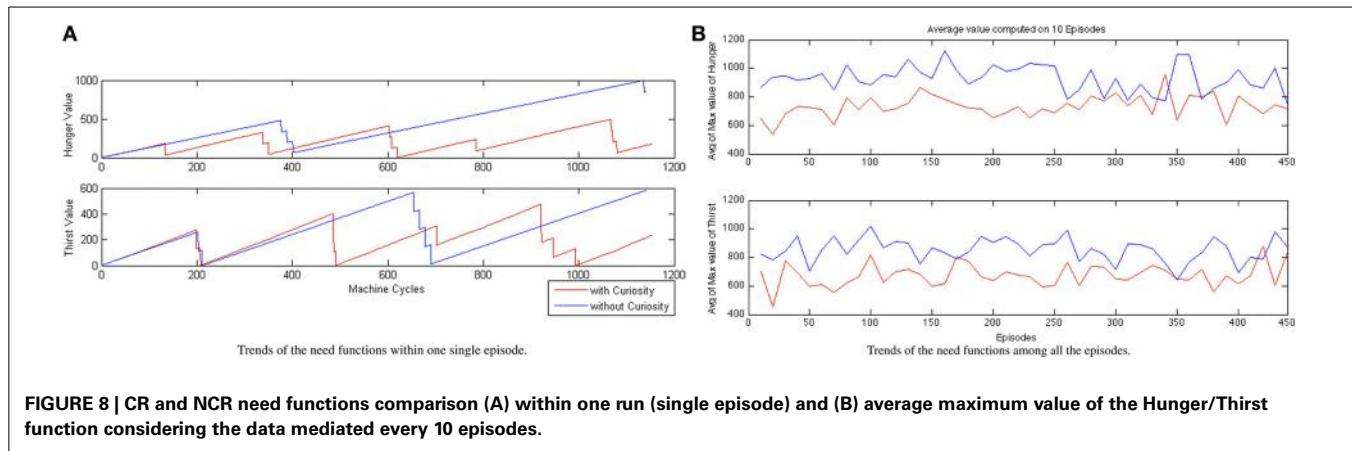
As we stated before, the robot must survive in the testing environment for a predefined amount of time (*max_time*). The plot in **Figure 6** shows that during the first 250 episodes the NCR system is more effective in surviving in the environment. In fact, the survival time percentage starts from a value over 90%. That is, the NCR system is more effective in action selection than the CR system. This could be due to the fact that the curiosity, initially, leads the system to prefer the exploration of novel spaces rather than the goal-directed ones. However, after a while, the CR system starts to rapidly increase its survival time until it over pass the NCR system and reaches the convergence (100% of the survival time) around the episode 300, with respect to the NCR system that does not reach the convergence before 450 episodes. Hence, in both the cases we observe that the learning converges after at most 450 episodes, however, in the case of the robot endowed with curiosity an earlier convergence is obtained.

In **Figure 7**, we show the cumulative rewards for each behavior during the learning process. The red lines describe the trend of the rewards gained by the system endowed with curiosity, while the blue ones are for the NCR system. As expected, during the first episodes the curious robot is not able to learn the attentional strategies needed to regulate the activations of the robot behaviors. The cumulative rewards related to the behavior AVOID of the CR system show that the performance remains unsatisfactory approximately until the episode 200. Then, the values of the cumulative rewards starts to increase and to converge from, approximately, the episode 300. In contrast, the NCR system shows a worst trend of the cumulative rewards for the EAT and DRINK behaviors. This could be explained by the fact that the robot is not guided by the curiosity to immediately explore the spaces of the environment where food or water are not observed. It only learns to eat or drink when the associated need functions (hunger and thirst) exceed a certain threshold; while the associated behaviors remains always relaxed. That is, the learned policy

for the NRC DRINK behavior always selects the maximum value for the period ($p^{drink} = 12$) for all the states associated to low levels of thirst (i.e., from σ_1 to σ_4). On the contrary, it selects the shorter period value ($p^{drink} = 1$) for the states with a high level of thirst (σ_5 and σ_6). In the case of the CR robot, the process of learning is affected by the curiosity, which influences the robot behavior to explore spaces of the environment with food or water sources since it is immediately attracted by novel stimuli (including green and blue blob). Hence, the CR system learns to eat or drink also when this is not strictly required. For example, the learned policy for the CR DRINK behavior, in the case of low curiosity, associates the maximum value for p^{drink} only to fewer states with low levels of thirst (from σ_1 to σ_3), and it selects short period values ($p^{drink} = 1$) for all the other states (from σ_4 to σ_6). Finally, the learned policy for the CR DRINK behavior, in the case of high curiosity, always associates $p^{drink} = 1$ or $p^{drink} = 2$ to all the levels of thirst (from σ_1 to σ_6). At the end of the experiments NCR EAT and DRINK rewards converge to higher values, however, the global reward is higher for the CR. The global cumulative rewards are collected in the last plot of **Figure 7** (on the bottom row), which shows the faster convergence of the CR system with respect to the NCR one. Finally, the CR learned policies for EAT and DRINK can always maintain the Energy value above the wellness threshold (this is also visible in **Figure 10**).

In **Figure 8A**, we show, respectively, (A) the trends of the need functions within a single episode after the convergence of the learning process, and (B) the trend of the average value of the maximum value of the need functions among all the episodes. If we look at their trends, in **Figure 8A** we observe some periodical path for the values of each function. We interpret the plots, in the case of the CR, as an effective learned attentional shifting policy of the behaviors EAT and DRINK. The robot seems to find a rhythmic alternation of its needs of eating and drinking (the decreasing part of the hunger and thirst functions corresponds



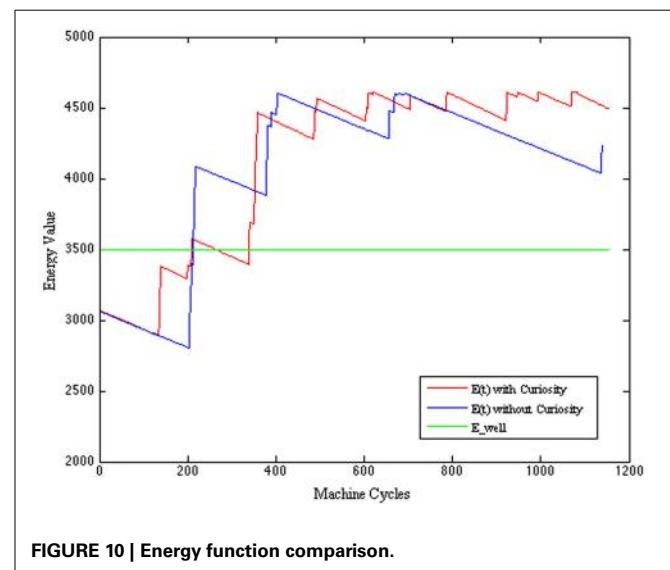


to the consuming of food or water, respectively). On the contrary, the NCR system just waits to become very hungry or very thirsty before starting to search for sources of food and water. The behavior of the NCR robot, while on the one hand, driven by the thirst and the hunger need functions, achieves better results in terms of the single rewards, on the other hand, this does not lead to a global better reward for the NCR with respect to the CR (see Figure 7). This is also visible in Figure 8B where we can observe that the need to eat or drink for the NCR is, on average, always greater than the CR needs. Thus, the CR system is able to find a configuration of the activation periods (i.e., suitable attentional monitoring strategies), associated with the EAT and DRINK behaviors, such that the robot never suffers because of some internal need, leading to a best homeostatic regulation of the internal variables.

In order to evaluate the performance of the two robots we defines a measure of safety as follows:

$$Safety(t) = \frac{\sigma_t^{avoid}}{\sigma_{max}^{avoid}} \cdot \frac{P_{max}^{avoid} - P_t^{avoid}}{P_{max}^{avoid} - P_{min}^{avoid}} \quad (13)$$

where the level of safety is calculated with respect to the minimum distance between the current position of the robot and an obstacle. Here, the danger increases when the distance decreases and the AVOID activation period is relaxed; and, viceversa, the safety increases when the activation period of the AVOID is suitably balanced with respect to the distance from an obstacle. The improved performance of the CR system is visible in the evaluation of the safety function (see Figure 9), where we observe more pleasurable values for the CR robot, and of the energy function (see Figure 10), where the CR system is able to maintain the levels



of energy E_c with Curiosity not only above the threshold of well-being E_{well} , but also stabilized at a high value. The eighth row of Table 5 shows the average rewards of AVOID, EAT, and DRINK behaviors and the averages values of the global reward for the 100 episodes of validation.

All the results of the above plots are summarized in Table 5, where we evaluate the average values and the standard deviations on 100 episodes used to validate our system (after the convergence of the learning process). Regarding the global energy of the system, we already noticed that such values stabilized on a specific interval for the CR (see Figure 10). In Table 5, we can find that the average of the Energy mean values ($= 3720$) is a bit smaller than the NCR case ($= 3809$) and that its maximum value ($= 4488$, which is above the wellness threshold E_{well}) is smaller than the NCR energy maximum value ($= 4525$). However, we suppose that the CR average of the energy mean values is smaller because of the curiosity (residual energy), which is “consumed” for exploring new states during the learning process. Interestingly, such an exploration of new states does not imply that the robot is less cautious in moving around. Indeed, the safety average value of the

Table 5 | Maximum, minimum and average values for the need functions (safety, hunger and thirst) and the energy function.

	Robot without Curiosity				Robot with Curiosity			
	Max		Min		Max		Min	
	Energy	Safety	Hunger	Thirst	Energy	Safety	Hunger	Thirst
Energy	4525 ± 247		2683 ± 212		4484 ± 344		2781 ± 115	
Safety	0.97 ± 0.09		0.12 ± 0.02		1.00		0.18 ± 0.04	
Average value	3809 ± 335	0.38 ± 0.02	420 ± 181	363 ± 135	3720 ± 312	0.61 ± 0.03	328 ± 142	311 ± 137
	Avoid	Eat	Drink	Global	Avoid	Eat	Drink	Global
Reward	66 ± 2	83 ± 6	85 ± 5	235 ± 7	100 ± 9	70 ± 7	78 ± 11	248 ± 15
Activation	109 ± 3	105 ± 7	103 ± 2	260 ± 22	204 ± 19	130 ± 10	156 ± 10	418 ± 22

Average values of the cumulative eat and drink rewards and of number of activations of the behaviors after the learning process.

CR is almost two times greater ($= 0.61$) than the NCR ($= 0.38$). Moreover, the minimum value for the CR safety ($= 0.18$) is higher (so, more safer) than the NCR value ($= 0.12$). Finally, as noted in the plots (see **Figure 8A**) the need functions of hunger and thirst have smaller average values for the CR (hunger = 328 and thirst = 311), which means that the robot satisfies its needs more frequently.

Both CR and NCR are effective in spending computational resources. This can be observed by considering the last row of **Table 5**, where we show the average number of the behavior activations. The CR has a slightly greater number of activations, in particular for the AVOID behavior. This leads to an emergent behavior consistent with what discussed above. The CR robot eats and drinks more frequently and shows a safer behavior with respect to the NCR. However, 204 activations out of *max_cycles* (around 1150 in this specific case) possible activations (machine cycles of an episode) seems a satisfactory result for a behavior-based architecture (i.e., there is a reduction of the 83% of the number of activations).

Moreover, notice that the global value shown at the end of this row states for how many cycles at least one behavior was active during the episode. In the case of NCR, this value is equal to 260. This shows that it is frequent to find more than one behavior active at the same time. For the CR robot, this value is equal to 418, meaning that for the most of the time only one behavior is active and the robot is able to orchestrate the multiple behaviors by opportunely shifting attentional resources, from time to time, toward the most salient one according to its need functions.

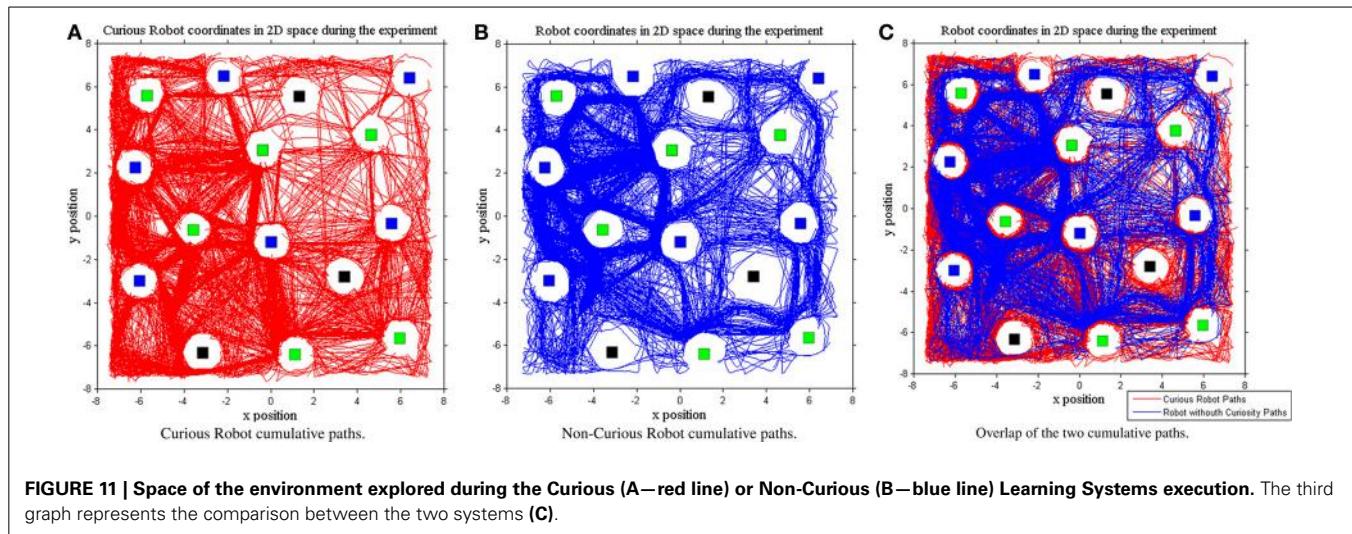
Finally, another interesting result regards the curiosity influence on the actual environmental exploration space. Indeed, while we expected that our intrinsic motivated RL would lead the learning process to improve the exploration of the internal learning states, we did not expect that this would also produce an increased spatial exploration of the environment. This result can be illustrated by plotting the paths of the two systems during the overall experimentation (see **Figure 11C**). By comparing the two generated paths, we can note that the system endowed with internal motivation (see **Figure 11A**) is more explorative (the cumulative traces of 500 episodes covered the 50% of the total area) with respect to the non-curios one (44% of the total area covered as shown in **Figure 11B**). The CR path seems smoother with a better coverage of the

space around obstacles, food, and water while keeping the robot safe.

DISCUSSION

In this paper, we presented an intrinsically motivated RL approach to attention allocation and shifting in a robotic system. The framework has been demonstrated at work in a survival domain. Differently from classical RL models of action selection, where actions are chosen according to the operative/perceptive contexts, in our case the action selection is mediated by the attentional status of the robotic behaviors. In the literature we can find intrinsically motivated RL system where simple attentional control mechanisms are involved (e.g., eye movement in the playground domain in Barto et al., 2004); in this paper we tackle the attention allocation and shifting problem, which is novel in this context. Indeed, in our setting, the learning process is to adapt and modulate the attentional strategies used to allocate attentional resources of the system. Specifically, our attentional mechanism regulates the behavioral activation periods, hence the amount of computational and operation resources dedicated to monitor and control the associated activities. Following this approach, the global behavior of the system is not directly generated by an action selection policy (as in typical RL approaches to action selection Sutton and Barto, 1998 and intrinsically motivated RL Barto et al., 2004; Singh et al., 2004), instead, it emerges as the sum of the outputs of multiple parallel processes, each activated with its own frequency: the smaller the activation period of a behavior, the higher its influence on the global emergent behavior. Following the taxonomy proposed by Baldassarre and Mirolli (2013), our system can also be considered as a competence-based system where the skill to be learned is the attentional allocation policy, however, this policy has only an indirect effect on the overall expected reward.

As the main intrinsic motivation, we considered the curiosity drive which is inspired by the one proposed by Litman (2005). This model allows us to account for both optimal arousal and curiosity-driven approaches to curiosity modeling. In particular, we related the liking and wanting drives of the Litman's model to, respectively, the pleasure of the novelty and the residual energy of the system (the higher the energy value over the wellness state, the higher the drive toward to the exploration of novel situations and states). While several models for



novelty-based and knowledge-based (Schmidhuber, 1991; Singh et al., 2004) curiosity have been proposed in the intrinsically motivated RL literature, the employment of the Litman account is less explored. Notice that we do not employ knowledge-based curiosity models (Schmidhuber, 1991; Singh et al., 2004). Indeed, while in Schmidhuber (1991) and later in Singh et al. (2004) and Oudeyer and Kaplan (2007) curiosity should lead the agent to explore areas of the environment where the learning progress is expected to be high, in our system, the agent is directly attracted by novel stimuli as sources of saliency. We want to stress here that the attentional problem addressed in our work is different from the ones mentioned since we learn attentional allocation only. In contrast to Schmidhuber (1991), Singh et al. (2004), and Oudeyer and Kaplan (2007), we can only enhance attention with respect to the attracting stimuli, but the movement of the system toward the stimuli is obtained as an indirect effect. As for the novelty, the lapsing mechanism we defined for the liking function (the novelty of a state is reduced when it is frequently visited within the time window) can be related to the Itti's model of surprise (Baldi and Itti, 2010), but also to the approach proposed by Oudeyer and Kaplan (2007), where, once predictions within a given part of the sensorimotor space are learned, the system gets bored and starts to execute other actions. As far as attentional allocation and shifting is concerned, RL models have been mainly proposed for visual attentions and gaze control (Bandera et al., 1996), a theoretical link between visual attentional exploration and novelty-based intrinsic motivations is investigated in Schlesinger (2012) where the author investigates the way in which goal directed, top-down attentional skills can be incrementally learned exploiting complex novelty detection strategies. Differently from these cases, here we investigated an intrinsically motivated RL approach to the generation of attentional strategies that are suitable for the executive control.

Our approach has been illustrated and tested in a simulated survival domain, where a robot was engaged in survival tasks such as finding food or water while avoiding dangerous situations. In this context, our goal was to show the feasibility and the

effectiveness of the approach in a typical robotic domain where basic needs satisfaction and intrinsic (curiosity) motivations were clearly defined. In particular, we compared the performance of the intrinsically motivated RL with respect to the same setting except for the fact that the influence of the intrinsic drive was neglected. The parameter tuning was provided in order to find the best regulation of the non-curious setting to assess the added value of the curiosity drive. The collected results support the hypothesis that the curiosity-driven learning system permits to find satisfactory regulations of the attention allocation and shifting policies, providing different attentional policies, and consequently different emergent behaviors, depending on the current level of curiosity. Moreover, the overall behavior that emerges from the execution of the learned attentional policies seems safer and capable of keeping the robotic system in a higher wellness state during the environment exploration. This is related to the fact that the curiosity drive stimulates the attention toward opportunities of energy recharging (food and water) more frequently than in the non-curious system. We also observed that the curious system provides a more uniform exploration of the environment when compared with the non-curious behavior. While the presented tests illustrate the feasibility and effectiveness of the approach in a typical survival domain, the extension of this curiosity-driven attentional regulation method to more complex domains and more structured tasks (e.g., considering hierarchical skills Barto et al., 2004 and top-down attentional regulations Schlesinger, 2012) remains to be investigated as future work.

ACKNOWLEDGMENTS

The research leading to these results has been supported by the SAPHARI Large-scale integrating project, which has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement ICT-287513. The authors are solely responsible for its content. It does not represent the opinion of the European Community and the Community is not responsible for any use that might be made of the information contained therein.

REFERENCES

- Arbib, M. A. (1998). "Schema theory," in *The Handbook of Brain Theory and Neural Networks*, ed M. A. Arbib (Cambridge, MA: MIT Press), 830–834.
- Arkin, R. C. (1998). *Behavior-Based Robotics*. Cambridge, MA: MIT Press.
- Baldassarre, G. (2011). "What are intrinsic motivations? A biological perspective," in *Proceedings of the IEEE Conference on Developmental Learning and Epigenetic Robotics* (Frankfurt am Main), 1–8.
- Baldassarre, G., and Mirolli, M. (2013). "Intrinsically motivated learning systems: an overview," in *Intrinsically Motivated Learning in Natural and Artificial Systems*, eds G. Baldassarre and M. Mirolli (Berlin: Springer-Verlag), 1–14. doi: 10.1007/978-3-642-32375-1_1
- Baldi, P. F., and Itti, L. (2010). Of bits and wows: a Bayesian theory of surprise with applications to attention. *Neural Netw.* 23, 649–666. doi: 10.1016/j.neunet.2009.12.007
- Bandera, C., Vico, F. J., Bravo, J. M., Harmon, M. E., and Iii, L. C. B. (1996). "Residual q-learning applied to visual attention," in *ICML-96* (Bari), 20–27.
- Baranes, A., and Oudeyer, P. Y. (2013). Active learning of inverse models with intrinsically motivated goal exploration in robots. *Robot. Auton. Syst.* 61, 49–73. doi: 10.1016/j.robot.2012.05.008
- Barto, A. G., Singh, S., and Chentanez, N. (2004). "Intrinsically motivated learning of hierarchical collections of skills," in *Proceedings of International Conference on Developmental Learning (ICDL)* (Cambridge, MA: MIT Press).
- Berlyne, D. E. (1954). A theory of human curiosity. *Br. J. Psychol.* 45, 180–191.
- Berlyne, D. E. (1960). *Conflict, Arousal and Curiosity*. New York, NY: McGraw-Hill. doi: 10.1037/11164-000
- Berridge, K. C. (2003). Pleasures of the brain. *Brain Cogn.* 52, 106–128. doi: 10.1016/S0278-2626(03)00014-9
- Brooks, R. (1986). A robust layered control system for a mobile robot. *IEEE J. Robot. Automat.* RA-2, 14–23. doi: 10.1109/JRA.1986.1087032
- Burattini, E., Finzi, A., Rossi, S., and Staffa, M. (2010). "Attentive monitoring strategies in a behavior-based robotic system: an evolutionary approach," in *Proceedings of the 2010 International Symposium on Learning and adaptive Behavior in Robotic System* (Canterbury: IEEE computer Society), 153–158.
- Burattini, E., and Rossi, S. (2008). Periodic adaptive activation of behaviors in robotic system. *Int. J. Pattern Recogn. Artif. Intell.* 22, 987–999. doi: 10.1142/S0218001408006661
- Dalley, J. W., Cardinal, R. N., and Robbins, T. W. (2004). Prefrontal executive and cognitive functions in rodents: neural and neurochemical substrates. *Neurosci. Biobehav. Rev.* 28, 771–784. doi: 10.1016/j.neubiorev.2004.09.006
- Di Nocera, D., Finzi, A., Rossi, S., and Staffa, M. (2012). "Attentional action selection using reinforcement learning," in *Proceedings of 12th International Conference on Adaptive Behaviour (SAB 2012)*, LNCS, Vol. 7426, eds T. Ziemke, C. Balkenius, and J. Hallam (Berlin: Springer), 371–380.
- Gerkey, B., Vaughan, R., and Howard, A. (2003). "The player/stage project: tools for multi-robot and distributed sensor systems," in *Proceedings of the International Conference on Advanced Robotics* (Coimbra), 317–323.
- Heidbreder, C. A., and Groenewegen, H. J. (2003). The medial prefrontal cortex in the rat: evidence for a dorso-ventral distinction based upon functional and anatomical characteristics. *Neurosci. Biobehav. Rev.* 27, 555–579. doi: 10.1016/j.neubiorev.2003.09.003
- Hull, C. L. (1943). *Principles of Behavior: An Introduction to Behavior Theory*. New York, NY: Appleton-Century-Croft.
- Itti, L., and Koch, C. (2001). Computational modeling of visual attention. *Nat. Rev. Neurosci.* 2, 194–203. doi: 10.1038/35058500
- Jepma, M., Verdonschot, R., van Steenbergen, H., Rombouts, S., and Nieuwenhuis, S. (2012). Neural mechanisms underlying the induction and relief of perceptual curiosity. *Front. Behav. Neurosci.* 6:5. doi: 10.3389/fnbeh.2012.00005
- Kaplan, F., and Oudeyer, P. Y. (2003). "Motivational principles for visual know-how development," in *Proceedings of the 3rd international workshop on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems*, eds C. G. Prince, L. Berthouze, H. Kozima, D. Bullock, G. Stojanov, and C. Balkenius (Lund: Lund University Cognitive Studies), 73–80.
- Lisman, J. E., and Grace, A. A. (2005). The hippocampal-vta loop: controlling the entry of information into long-term memory. *Neuron* 46, 703–713. doi: 10.1016/j.neuron.2005.05.002
- Litman, J. (2005). Curiosity and the pleasures of learning: wanting and liking new information. *Cogn. Emot.* 19, 793–814. doi: 10.1080/02699930541000101
- Loewenstein, G. (1994). The psychology of curiosity: a review and reinterpretation. *Psychol. Bull.* 75–98. doi: 10.1037/0033-2909.116.1.75
- Marsland, S., Nehmzow, U., and Shapiro, J. (2000). *A Real-Time Novelty Detector for a Mobile Robot*. CoRR cs.RO/0006006.
- Mirolli, M., and Baldassarre, G. (2013). "Functions and mechanisms of intrinsic motivations: the knowledge versus competence distinction," in *Intrinsically Motivated Learning in Natural and Artificial Systems*, eds G. Baldassarre and M. Mirolli (Berlin: Springer-Verlag), 49–72. doi: 10.1007/978-3-642-32375-1_3
- Montague, P. R., Dayan, P., and Sejnowski, T. K. (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *J. Neurosci.* 16, 1936–1947.
- Nieoullon, A. (2002). Dopamine and the regulation of cognition and attention. *Prog. Neurobiol.* 67, 53–83. doi: 10.1016/S0301-0082(02)00011-4
- Norman, D., and Shallice, T. (1986). Attention in action: willed and automatic control of behaviour. *Conscious. Self Regul. Adv. Res. Theor.* 4, 1–18. doi: 10.1007/978-1-4757-0629-1_1
- Oudeyer, P. Y., and Kaplan, F. (2006). What is intrinsic motivation? A typology of computational approaches. *Front. Neurorobot.* 1:6. doi: 10.3389/neuro.12.006.2007
- Oudeyer, P. Y., Kaplan, F., and Hafner, V. F. (2007). Intrinsic motivation systems for autonomous mental development. *IEEE Trans. Evol. Comput.* 11, 265–286. doi: 10.1109/TEVC.2006.890271
- Pashler, H., and Johnston, J. C. (1998). "Attentional limitations in dual-task performance," in *Attention*, ed H. Pashler (East Essex, UK: Psychology Press), 155–189.
- Posner, M., Snyder, C., and Davidson, B. (1980). Attention and the detection of signals. *J. Exp. Psychol. Gen.* 109, 160–174. doi: 10.1037/0096-3445.109.2.160
- Redgrave, P., and Gurney, K. (2006). The short-latency dopamine signal: a role in discovering novel actions? *Nat. Rev. Neurosci.* 7, 967–975. doi: 10.1038/nrn2022
- Ryan, R. M., and Deci, E. L. (2000). Intrinsic and extrinsic motivations: classic definitions and new directions. *Contemp. Educ. Psychol.* 25, 54–67. doi: 10.1006/ceps.1999.1020
- Schembri, M., Mirolli, M., and Baldassarre, G. (2007). "Evolving internal reinforcers for an intrinsically motivated reinforcement-learning robot," in *Proceedings of the 6th International Conference on Development and Learning*, eds Y. Demiris, D. Mareschal, B. Scassellati, and J. Weng (London: Imperial College), E1–E6.
- Schlesinger, M. (2012). "Investigating the origins of intrinsic motivations in human infants," in *Intrinsically Motivated Learning in Natural and Artificial Systems*, eds G. Baldassarre and M. Mirolli (Berlin: Springer-Verlag).
- Schmidhuber, J. (1991). "A possibility for implementing curiosity and boredom in model-building neural controllers," in *From Animals to Animats: Proceedings of the First International Conference on Simulation of Adaptive Behavior*, eds J.-A. Meyer and S. Wilson (Cambridge, MA: MIT Press), 222–227.
- Senders, J. (1964). The human operator as a monitor and controller of multidegree of freedom systems. *IEEE Trans. Hum. Fact. Electron.* HFE-5, 2–6. doi: 10.1109/THFE.1964.231647
- Sidobre, D., Broquere, W., Mainprice, J., Burattini, E., Finzi, A., Rossi, S., et al. (2012). "Human-robot interaction," in *Advanced Bimanual Manipulation*. Springer tracts in advanced robotics, Vol. 80, ed B. Siciliano (New York: Springer), 123–172. doi: 10.1007/978-3-642-29041-1_3
- Singh, S. P., Barto, A. G., and Chentanez, N. (2004). "Intrinsically motivated reinforcement learning," in *Advances in Neural Information Processing Systems 17: Proceedings of the 2004 Conference*, eds L. K. Saul, Y. Weiss, and L. Bottou (Cambridge, MA: MIT Press), 1281–1288.
- Singh, S., Lewis, R. L., Barto, A. G., and Sorg, J. (2010). Intrinsically motivated reinforcement learning: an evolutionary perspective. *IEEE Trans. Auton. Ment. Dev.* 2, 70–82. doi: 10.1109/TAMD.2010.2051031
- Spielberger, C. D., and Starr, L. M. (1994). "Curiosity and exploratory behavior," in *Motivation: Theory and Research*, eds H. F. O'Neil, Jr. and M. Drillings (Hillsdale, NJ: Erlbaum), 221–243.

- Sutton, R. and Barto, A. (1998). Reinforcement learning: an introduction, Vol. 1 (Cambridge, MA: Cambridge University Press).
- Watkins, C. and Dayan, P. (1992). Q-learning. *Mach. Learn.* 8, 279–292. doi: 10.1007/BF00992698
- White, R. (1959). Motivation reconsidered: the concept of competence. *Psychol. Rev.* 66, 297–333. doi: 10.1037/h0040934

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 21 June 2013; accepted: 13 March 2014; published online: 01 April 2014.

Citation: Di Nocera D, Finzi A, Rossi S and Staffa M (2014) The role of intrinsic motivations in attention allocation and shifting. *Front. Psychol.* 5:273. doi: 10.3389/fpsyg.2014.00273

This article was submitted to Cognitive Science, a section of the journal *Frontiers in Psychology*.

Copyright © 2014 Di Nocera, Finzi, Rossi and Staffa. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Novelty, attention, and challenges for developmental psychology

Emily Mather*

Department of Psychology, University of Hull, Hull, UK

*Correspondence: emily.mather@hull.ac.uk

Edited by:

Marco Mirolli, Istituto di Scienze e Tecnologie della Cognizione, Italy

INTRODUCTION

In this brief essay, I seek to demonstrate the significance of exploratory behavior for understanding cognitive development. Historically, organisms were thought to act solely in the service of achieving biologically significant goals, such as satisfying thirst, hunger, and reproductive drives. However, it became apparent that both animals and humans engage in behavior where the adaptive goal is unclear (see Hunt, 1963, 1965). With no obvious external target, this activity is best described as being intrinsically motivated, and often directed toward the unknown and the unexpected (Kagan, 2002). Hence *novelty*, the discrepancy between what is known and what is discovered, can elicit activity and exploration of the environment.

What is the relevance to developmental process? Attention to novelty plays a seemingly simple role in learning and development, directing the senses toward what is as yet unknown. Yet, research shows that patterns of attention to novelty are not straightforward, particularly during infancy. There is considerable evidence that attention is sometimes biased toward *familiarity*, rather than novelty. Unlike our understanding of novelty preference, we struggle to understand when and why familiarity preferences occur. Below I briefly review this area of research and illustrate how this basic aspect of learning continues to puzzle developmental psychologists.

FROM ANIMALS TO INFANTS?

The habituation mechanism, which directs attention to novelty, has been widely-studied across the animal kingdom (Sokolov, 1963; Thompson and Spencer, 1966). Thorpe (1963) defined habituation as “the relatively permanent waning of a response as a result of repeated

stimulation . . . ” (p. 61). An important feature of habituation is that an organism’s responding will recover to the presentation of a different stimulus—an effect known as dishabituation (Thompson and Spencer, 1966). Hence, habituation is stimulus-specific and attention will recover to novel stimuli. A seminal study by Fantz (1964) demonstrated that infants’ visual attention to a familiar, repeated image will decrease relative to their attention to a novel image. Other early studies of infant habituation also reveal infants’ interest in novel stimuli (see Cohen and Gelber, 1975, for a review). There is evidence that even newborns will habituate and direct their attention to novelty (Friedman, 1972; Slater et al., 1982, 1984). Many infancy researchers are interested in the use of novelty preferences as a methodological tool. Habituation-dishabituation procedures are used to demonstrate infants’ discrimination of stimuli, and seemingly precocious cognitive abilities (e.g., Baillargeon, 1987; Spelke et al., 1992; but see Schilling, 2000, for a counterargument).

Infants’ interest in novelty is consistent with theories of habituation accounting for both human and animal responding. However, there is substantial evidence that infants do not always prefer a novel stimulus—sometimes they prefer to attend to a familiar stimulus. In Rose et al. (1982) groups of 3.5- and 6.5-month-olds were exposed to a visual stimulus for different durations. This familiar stimulus was then paired with a novel stimulus. Infants at both ages displayed familiarity preferences after shorter exposures, and novelty preferences after longer exposures. Similarly, Hunter et al. (1983) presented 8- and 12-month-olds with a set of toys, and tested their preference for the familiar vs. novel toys after differing amounts of familiarization. The infants

preferred the familiar toys after a shorter familiarization period, and the novel toys after a longer familiarization period. In these studies, familiarization time was manipulated between groups of infants. Roder et al. (2000) provide a within-infants demonstration that 4.5-month-olds preferences shift from familiarity to novelty as a function of familiarization time.

While familiarity and novelty preferences have largely been investigated for the visual modality, there is also evidence that infants display these preferences for auditory stimuli (e.g., Colombo and Bundy, 1983; Spence, 1996). More recent research has found that infants’ preferences will also “reverse” as their memory for a familiarized stimulus decays over time. In Bahrick and Pickens (1995), 3-month-olds displayed a novelty preference after a 1 min retention interval, and a familiarity preference after a 1 month interval (see also Spence, 1996; Bahrick et al., 1997; Courage and Howe, 2001). Familiarity preferences do not just occur to repetitions of a specific stimulus. Infants who have categorized a set of stimuli will sometimes attend more to a novel stimulus from the same category, rather than a novel stimulus drawn from a novel category (e.g., Gomez and Gerken, 1999; Fiser and Aslin, 2002; Gómez and Maye, 2005; Mather and Plunkett, 2011). Hunter and Ames (1988) provide a descriptive model of infants’ familiarity and novelty preferences. The main factor is familiarization time—with briefer exposures, the infant attends more to a familiar stimulus, but with longer exposures, their attention turns to novelty. How quickly an infant makes this “familiarity-to-novelty shift” will depend on their processing speed and the complexity of the stimuli. Hence, familiarity preferences are more likely for younger

infants (slower processors), and for more complex stimuli.

Familiarity preferences are not consistent with the habituation process, where attention simply declines with repeated exposure to a stimulus (Thompson and Spencer, 1966). Some computational models of infant attention have tentatively linked familiarity preferences with the process of *sensitization* (Sirois and Mareschal, 2004; Schoner and Thelen, 2006). Sensitization occurs when the presentation of a stimulus leads to heightened behavioral responding. Importantly, if a stimulus is repeated, it can have the effect of sensitizing itself. Under dual-process theory (Groves and Thompson, 1970), habituation and sensitization are separate, opposing processes which interact to determine responding. Sensitization is related to stimuli intensity, and decays quickly. If sensitization is initially stronger than habituation, there will be an early increase in responding to a repeated stimulus, followed by a decrease. This pattern of response to a repeated stimulus is similar to the familiarity–novelty shift sometimes evidenced by infants. However, in contrast to habituation, sensitization will generalize to a wide range of stimuli (see Domjan, 1998, for examples). Hence, while sensitization can occur to a repeated stimulus, it would also generalize to other stimuli if they were also present. This means that sensitization cannot account for the stimulus specificity of familiarity and novelty preferences (see also Turk-Browne et al., 2008, for a related argument).

THE OLD OR THE NEW?

An alternative theoretical perspective could account for the existence of familiarity preferences. Since the 1950's, a variety of arguments have been made that both adults and infants prefer stimuli which provide an optimal level of novelty or information (Dember and Earl, 1957; Berlyne, 1960; McCall and McGhee, 1977). The optimum is defined by a "moderate" discrepancy between a stimulus and an observer's representation of that stimulus. Hence, the more discrepant a stimulus is from the observer's state of knowledge, the more novel it is to the observer. Relatedly, stimulus complexity influences the amount of learning required to reduce

this discrepancy. Any stimulus which is more or less discrepant than the optimum is of less interest to the observer. The familiarity-to-novelty shift displayed by infants is consistent with optimal-level theory. It is possible for a familiar stimulus to be favored over a novel stimulus, because the familiar stimulus could initially be closer to the optimum. Further processing of the familiar stimulus will result in a shift away from the optimum, and a novel stimulus will be preferred (see Hunter and Ames, 1988, for an elaboration).

A problem with obtaining evidence of the familiarity-to-novelty shift is that there is a temporally limited window for observing a familiarity preference. At a certain point, attention will shift toward novelty, thus a successful experimental design must be sensitive to this shift. Different infants will also process information at different rates, meaning that individual preferences can be obscured by group data (see Roder et al., 2000). Unfortunately, optimal-level theory does not provide a remedy for these methodological issues. The key variables involved—stimulus complexity, processing speed, and the optimal level of novelty—are usually unknown quantities. This makes it difficult to predict the occurrence of familiarity preferences, and to test the assumptions of the theory (see Thomas, 1971). A lack of familiarity preference could be due to the stimuli not being sufficiently complex, or an infant rapidly processing the stimuli. Therefore, while the existence of familiarity preferences is consistent with optimal-level theory, the theory itself perhaps does little more than assert that we seek out moderately novel stimuli.

One approach to dealing with the shortcomings of optimal-level theory has been to develop more computationally explicit models of familiarity and novelty preferences (Sirois and Mareschal, 2004; Schoner and Thelen, 2006; see also Perone et al., 2011) and to mathematically formalize the information content of a stimulus (e.g., Kidd et al., 2012). These recent advances offer an improved level of theoretical precision over past formulations of optimal-level theory. Nonetheless, these models incorporate some of the basic assumptions of optimal-level theory, and may also retain the difficulties

of predicting the familiarity-to-novelty shift. Our ability to understand exactly when infants will seek out familiarity or novelty is likely to require a deeper understanding of *why* there is an optimum in the first place. Development requires a balance of familiarization with regularities in the environment (Gibson, 1969) and shifting attention to what is new and unknown so as to create new cognitive structures (Piaget, 1936/1952). Therefore, rather than just focusing on preferences for individual stimuli, one useful approach might be to explore the more global consequences for the abstraction and development of knowledge.

ORDER AND TIMING: THE CYCLE OF COGNITIVE DEVELOPMENT

The familiarity-to-novelty shift causes infants to process stimuli in a particular sequence. That is, with all other factors held constant, infants' will explore different stimuli in a systematic fashion, based on their prior experience and learning. Beyond the laboratory, how do these preferences shape patterns of learning across the vast multitude of items and events in the real world, across multiple timescales? Computational models and experimental data demonstrate how the pattern of input can influence the trajectory and success of learning. In Elman (1993), recurrent neural networks were more successful at acquiring grammatical categories if they began by only learning about a subset of the total sentences available, rather than learning about all sentences together (see also Plunkett and Marchman, 1991, 1993). Other research suggests that order effects may also occur in infant categorization (Sandhofer and Doumas, 2008; Mather and Plunkett, 2011). What is particularly intriguing is that in some cases, exposure to an initially restricted stimulus set supports learning (Elman, 1993), whereas in other cases, reduced variability hinders learning (see Mather and Plunkett, 2011). These findings hint at a global effect of optimal preferences on successful learning.

Currently, we understand little about the role that familiarity and novelty preferences might play in driving successful patterns of learning. However, if we can better understand the effects of these preferences

on cognitive development, then we might make sense of the underlying cause of optimal preferences. Conversely, our explanations of cognitive development would also benefit from understand the impact of exploratory behavior on learning. Much current developmental research is concerned with specifying the mechanisms of learning, without considering how and why attention prioritizes certain stimuli for learning. Cognitive development needs to be understood as a cyclical process, where attention influences learning, and learning guides attention. If, as Piaget (1936/1952) argued, the child is *actively engaged in the construction of their own knowledge*, then exploratory behavior needs to be placed at the heart of cognitive development.

CONCLUSIONS

A hallmark of human behavior is that we seek to explore and understand our environments, even in the absence of biological or externally specified goals. Our interest in what is new and unknown is evident from birth. However, we do not yet have a clear understanding of the mechanisms which determine whether a child attends to familiarity or novelty. The function of optimal preferences may need to be interpreted in the context of broader developmental changes. Both the processes of cognitive growth and exploratory behavior can be better understood by considering their interdependence.

REFERENCES

- Bahrick, L. E., Hernandez-Reif, M., and Pickens, J. N. (1997). The effect of retrieval cues on visual preferences and memory in infancy: evidence for a four-phase attention function. *J. Exp. Child Psychol.* 67, 1–20. doi: 10.1006/jecp.1997.2399
- Bahrick, L. E., and Pickens, J. N. (1995). Infant memory for object motion across a period of three months: implications for a four-phase attention function. *J. Exp. Child Psychol.* 59, 343–371. doi: 10.1006/jecp.1995.1017
- Baillargeon, R. (1987). Object permanance in 3.5- and 4.5-month-old infants. *Dev. Psychol.* 23, 655–664. doi: 10.1037/0012-1649.23.5.655
- Berlyne, D. E. (1960). *Conflict, Arousal, and Curiosity*. New York, NY: McGraw-Hill. doi: 10.1037/11164-000
- Cohen, L. B., and Gelber, E. R. (1975). “Infant visual memory,” in *Infant Perception: From Sensation to Cognition*, Vol. 1, eds L. B. Cohen and P. Salapatek (London: Academic Press), 347–403.
- Colombo, J., and Bundy, R. S. (1983). Infant response to auditory familiarity and novelty. *Infant Behav. Dev.* 6, 305–311. doi: 10.1016/S0163-6383(83)80039-3
- Courage, M. L., and Howe, M. L. (2001). Long-term retention in 3.5-month-olds: familiarization time and individual differences in attentional style. *J. Exp. Child Psychol.* 79, 271–293. doi: 10.1006/jecp.2000.2606
- Dember, W. N., and Earl, R. W. (1957). Analysis of exploratory, manipulatory, and curiosity behaviors. *Psychol. Rev.* 64, 91–96. doi: 10.1037/h0046861
- Domjan, M. (1998). *The Principles of Learning and Behaviour*. 4th Edn. Pacific Grove, CA: Brooks/Cole.
- Elman, J. L. (1993). Learning and development in neural networks: the importance of starting small. *Cognition* 48, 71–99. doi: 10.1016/0010-0277(93)90058-4
- Fantz, R. L. (1964). Visual experience in infants: decreased attention to familiar patterns relative to novel ones. *Science* 146, 668–670. doi: 10.1126/science.146.3644.668
- Fiser, J., and Aslin, R. N. (2002). Statistical learning of new visual feature combinations by infants. *Proc. Natl. Acad. Sci. U.S.A.* 99, 15822–15826. doi: 10.1073/pnas.232472899
- Friedman, S. (1972). Habituation and recovery of visual response in the alert human newborn. *J. Exp. Child Psychol.* 13, 339–349. doi: 10.1016/0022-0965(72)90095-1
- Gibson, E. (1969). *Principles of Perceptual Learning and Development*. New York, NY: Appleton-Century-Crofts.
- Gomez, R. L., and Gerken, L. (1999). Artificial grammar learning by 1-year-olds leads to specific and abstract knowledge. *Cognition* 70, 109–135. doi: 10.1016/S0010-0277(99)00003-7
- Gómez, R. L., and Maye, J. (2005). The developmental trajectory of nonadjacent dependency learning. *Infancy* 7, 183–206. doi: 10.1207/s15327078inf0702_4
- Groves, P. M., and Thompson, R. F. (1970). Habituation: a dual-process theory. *Psychol. Rev.* 77, 419–450. doi: 10.1037/h0029810
- Hunt, J. McV. (1963). “Motivation inherent in information processing and action,” in *Motivation and Social Interaction: Cognitive Determinants*, ed O. J. Harvey (New York, NY: Ronald Press), 35–94.
- Hunt, J. McV. (1965). “Intrinsic motivation and its role in psychological development,” in *Nebraska Symposium on Motivation*, Vol. 13, ed D. Levine (Lincoln, NB: University of Nebraska), 189–282.
- Hunter, M. A., and Ames, E. W. (1988). “A multifactor model of infant preferences for novel and familiar stimuli,” in *Advances in Infancy Research*, Vol. 5, eds C. Rovee-Collier and L. P. Lipsitt (Stamford, CT: Ablex), 69–95.
- Hunter, M. A., Ames, E. W., and Koopman, R. (1983). Effects of stimulus complexity and familiarisation time on infant preferences for novel and familiar stimuli. *Dev. Psychol.* 19, 338–352. doi: 10.1037/0012-1649.19.3.338
- Kagan, J. (2002). *Surprise, Uncertainty, and Mental Structures*. Cambridge, MA: Harvard University Press.
- Kidd, C., Piantadosi, S. T., and Aslin, R. N. (2012). The Goldilocks effect: human infants allocate attention to visual sequences that are neither too simple nor too complex. *PLoS ONE* 7:e36399. doi: 10.1371/journal.pone.0036399
- Mather, E., and Plunkett, K. (2011). Same items, different order: effects of temporal variability on infant categorization. *Cognition* 119, 438–447. doi: 10.1016/j.cognition.2011.02.008
- McCall, R. B., and McGhee, P. E. (1977). “The discrepancy hypothesis of attention and affect in infants,” in *The Structuring of Experience*, eds I. C. Uzgiris and F. Weizmann (New York, NY: Plenum), 179–210.
- Perone, S., Simmering, V. R., and Spencer, J. P. (2011). Stronger neural dynamics capture changes in infants’ visual working memory capacity over development. *Dev. Sci.* 14, 1379–1392. doi: 10.1111/j.1467-7687.2011.01083.x
- Piaget, J. (1936/1952). *The Origins of Intelligence in Children*. New York, NY: International Universities Press.
- Plunkett, K., and Marchman, V. (1991). U-shaped learning and frequency effects in a multi-layered perceptron: implications for child language acquisition. *Cognition* 38, 43–102. doi: 10.1016/0010-0277(91)90022-V
- Plunkett, K., and Marchman, V. (1993). From rote learning to system building: acquiring verb morphology in children and connectionist nets. *Cognition* 48, 21–69. doi: 10.1016/0010-0277(93)90057-3
- Roder, B. J., Bushnell, E. W., and Saserville, A. M. (2000). Infants’ preferences for familiarity and novelty during the course of visual processing. *Infancy* 1, 491–507. doi: 10.1207/S15327078IN0104_9
- Rose, S. A., Melloy-Carminar, P., Gottfried, A. W., and Bridger, W. H. (1982). Familiarity and novelty preferences in infant recognition memory: implications for information processing. *Dev. Psychol.* 18, 704–713. doi: 10.1037/0012-1649.18.5.704
- Sandhofer, C. M., and Doumas, L. A. A. (2008). Order of presentation effects in learning color categories. *J. Cogn. Dev.* 9, 194–221. doi: 10.1080/15248370802022639
- Schilling, T. H. (2000). Infants’ looking at possible and impossible screen rotations: the role of familiarization. *Infancy* 1, 389–402. doi: 10.1207/S15327078IN0104_2
- Schoner, G., and Thelen, E. (2006). Using dynamic field theory to rethink infant habituation. *Psychol. Rev.* 113, 273–299. doi: 10.1037/0033-295X.113.2.273
- Sirois, S., and Mareschal, D. (2004). An interacting systems model of infant habituation. *J. Cogn. Neurosci.* 16, 1352–1362. doi: 10.1162/0898929042304778
- Slater, A., Morison, V., and Rose, D. (1982). Visual memory at birth. *Br. J. Psychol.* 73, 519–525. doi: 10.1111/j.2044-8295.1982.tb01834.x
- Slater, A., Morison, V., and Rose, D. (1984). Habituation in the newborn. *Infant Behav. Dev.* 7, 183–200. doi: 10.1016/S0163-6383(84)80057-0
- Sokolov, E. N. (1963). *Perception and the Conditioned Reflex*. Oxford: Pergamon Press.
- Spelke, E. S., Breinlinger, K., Macomber, J., and Jacobson, K. (1992). Origins of knowledge.

- Psychol. Rev.* 99, 605–632. doi: 10.1037/0033-295X.99.4.605
- Spence, M. J. (1996). Young infants' long-term auditory memory: evidence for changes in preference as a function of delay. *Dev. Psychobiol.* 29, 685–695.
- Thomas, H. (1971). Discrepancy hypotheses: methodological and theoretical considerations. *Psychol. Rev.* 78, 249–259. doi: 10.1037/h0030795
- Thompson, R. F., and Spencer, W. A. (1966). Habituation: a model phenomenon for the study of neuronal substrates of behavior. *Psychol. Rev.* 73, 16–43. doi: 10.1037/h0022681

- Thorpe, W. H. (1963). *Learning and Instinct in Animals, 2nd Edn.* London: Methuen.
- Turk-Browne, N. B., Scholl, B. J., and Chun, M. M. (2008). Babies and brains: habituation in infant cognition and functional neuroimaging. *Front. Hum. Neurosci.* 2:16. doi: 10.3389/neuro.09.016.2008

Received: 20 May 2013; accepted: 13 July 2013; published online: 01 August 2013.

Citation: Mather E (2013) Novelty, attention, and challenges for developmental psychology. *Front. Psychol.* 4:491. doi: 10.3389/fpsyg.2013.00491

This article was submitted to Frontiers in Cognitive Science, a specialty of Frontiers in Psychology.

Copyright © 2013 Mather. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Autonomous visual exploration creates developmental change in familiarity and novelty seeking behaviors

Sammy Perone * and John P. Spencer

Department of Psychology and Delta Center, University of Iowa, Iowa City, IA, USA

Edited by:

Tom Stafford, University of Sheffield, UK

Reviewed by:

Matthew Schlesinger, Southern Illinois University, USA

Emily Mather, University of Hull, UK

***Correspondence:**

Sammy Perone, Department of Psychology and Delta Center, University of Iowa, E11 Seashore Hall, Iowa City, 52242 IA, USA
e-mail: sammy-perone@uiowa.edu

What motivates children to radically transform themselves during early development? We addressed this question in the domain of infant visual exploration. Over the first year, infants' exploration shifts from familiarity to novelty seeking. This shift is delayed in preterm relative to term infants and is stable within individuals over the course of the first year. Laboratory tasks have shed light on the nature of this familiarity-to-novelty shift, but it is not clear what motivates the infant to change her exploratory style. We probed this by letting a Dynamic Neural Field (DNF) model of visual exploration develop itself via accumulating experience in a virtual world. We then situated it in a canonical laboratory task. Much like infants, the model exhibited a familiarity-to-novelty shift. When we manipulated the initial conditions of the model, the model's performance was developmentally delayed much like preterm infants. This delay was overcome by enhancing the model's experience during development. We also found that the model's performance was stable at the level of the individual. Our simulations indicate that novelty seeking emerges with no explicit motivational source via the accumulation of visual experience within a complex, dynamical exploratory system.

Keywords: visual exploration, dynamic systems, dynamic neural fields, intrinsic motivation

One of the oldest questions in the history of human thought is what motivates an individual to achieve a level just beyond reach. Such motivation appears to be a central quality of human behavior, and may be a driving force behind scientific advancement, corporate innovation, and more generally, cultural evolution. Striving beyond one's reach is also an apt characterization of human development where children undergo a series of astonishing transformations. The newborn has a limited repertoire including sleeping, eating, and crying. By the end of the first year, the infant can walk and is beginning to talk. By age 5, the child is learning to read, write, and sit in a classroom among peers. What motivates a child to accomplish so much in so little time?

Seminal theories of cognitive development posit that infants' active exploration of their environments enables them to develop skilled action and cognitive systems (Piaget, 1952; Gibson, 1988). Infants are seemingly driven to act by curiosity, ambiguity, and novelty. These forces characterize intrinsic motivation and are widely held in developmental psychology to propel development forward (for a review, see Oudeyer and Kaplan, 2007). Yet the nature of intrinsic motivation and the mechanisms by which it creates change remain unclear.

Infancy might offer unique insights into the very nature of intrinsic motivation and its role in development. But how do we investigate intrinsic motivation in infants who have a limited behavioral repertoire? This requires clever methods to assess how infants think. Such methods first emerged in the 1970s when researchers developed a battery of novel habituation paradigms that relied on infants' looking behavior to measure cognition (Cohen, 1972a,b; Fantz, 1974). In these paradigms, infants are given experience looking at one item in isolation or in pairs. Then,

infants' preference to look at a novel item relative to the familiar item is measured. Infants' preference to look at a familiar over a novel item is taken as evidence that they recognize the familiar item but have not yet formed a robust memory for it. Infants' preference for novelty is taken as evidence that they have formed a robust memory for the familiar item and are beginning to learn about the novel item.

The use of looking paradigms led to the accumulation of a vast literature on infant cognition. A key finding from this literature is that infants' familiarity and novelty preferences change across multiple timescales, including during learning within a task and over weeks, months, and years in development (for reviews, see Hunter and Ames, 1988; Rose et al., 2004). With only brief exposure to a stimulus, infants will exhibit a familiarity preference. After prolonged exposure to the stimulus, infants will exhibit a novelty preference (Rose et al., 1982; for exceptions and detailed analysis, see Roder et al., 2000; Fisher-Thompson and Peterson, 2004). Critically, the rate at which infants move through this familiarity-to-novelty shift increases with age. In fact, during the first 1–2 months of life, infants move through this shift so slowly that they sometimes show no novelty preference even after several minutes of exposure (Wetherford and Cohen, 1973; Fantz, 1974). With age, however, infants spend more time looking at novel items relative to familiar items (Fantz, 1974).

This characterization of familiarity and novelty preferences is, of course, somewhat oversimplified. Infants' preferences are influenced by stimulus conditions, for instance (for reviews, see Hunter and Ames, 1988; Rose et al., 2004). For some stimuli, infants show no evidence of familiarity preferences early in learning (Roder et al., 2000). For other stimuli, infants show a

familiarity preference late in learning (Shinskey and Munakata, 2005). To complicate matters further, some studies have shown that individual infants oscillate between familiarity and novelty as they explore items (Fisher-Thompson and Peterson, 2004). And even adults will show familiarity preferences under conditions in which they freely explore visual scenes (Dodd et al., 2009). Thus, the same exploratory system appears to organize itself differently across contexts.

In the present report, we focus on the robust, quantitative increase in infants' exploration of novelty over development (for a broader theoretical evaluation of the familiarity-to-novelty shift, see Perone and Spencer, 2013a). This shift has been attributed to an increase in visual processing speed over development. Rose et al. (2002) nicely quantified this shift using a processing speed task with 5-, 7-, and 12-month-old infants. Infants were presented with pairs of different stimuli across trials. On each trial, one stimulus remained unchanged (familiar) and one changed (novel). This design enabled Rose et al. to quantify the time infants' spent looking at the familiar item before shifting over to explore novel items. Processing speed was measured as the number of trials to a criterion defined as a looking preference for the novel item on three consecutive trials. With age, infants accumulated less time looking to the familiar item and more quickly shifted toward looking to the novel item. This resulted in a reduction in the number of trials to reach criterion over development.

The use of looking paradigms has also led to two other key observations. First, infants' birth status influences the development of the familiarity-to-novelty shift. For example, Rose et al. (2002) found that term and preterm infants exhibited different patterns of familiarity and novelty seeking over development. At each age group, preterm infants required more trials to criterion than term infants. Thus, preterm infants exhibited stronger familiarity seeking biases than term infants and those persisted over development. Second, individual differences in looking behavior during infancy are stable over time. For example, Rose et al. (2001; see also Colombo et al., 1987) found that looking measures of exploration (e.g., frequency of gaze switching) and recognition (e.g., preference for novelty) are stable within individuals over the course of the first year. In addition, these looking measures during infancy are predictive of cognition during toddlerhood (Rose et al., 2009) and children's executive functioning at age 11 (Rose et al., 2012).

These laboratory-based observations have shed important light on the nature of the transition from familiarity- to novelty-seeking in the first year. Novelty-seeking has some distinct advantages over familiarity-seeking, enabling infants to explore and acquire knowledge about new items. Moreover, this exploratory process builds a strong base of what is familiar to the infant. But it is not clear from these data what motivates infants to switch their exploratory style. Conceptual and formal theories of infant looking and memory formation have attributed this shift to increases in processing speed (for reviews, see Hunter and Ames, 1988; Rose et al., 2004). By this view, infants' switch in exploratory style is simply a by-product of more efficient processing of visual information in the neural systems involved in doing so (Colombo,

1995). Although compelling, such accounts rarely explain where changes in processing speed come from.

Insights into this question might be obtained by moving from constrained laboratory tasks to less constrained tasks where infants can freely and autonomously explore the world around them. A nice example of this comes from recent studies of the transition from crawling to walking. What motivates an infant to move from skilled crawling to unskilled walking? Why move from an energy-efficient strategy to an energy-inefficient strategy? Adolph et al. (2012; see also Adolph and Robinson, 2013) observed infants' who were learning to walk in more naturalistic settings and made two surprising observations. First, infants engage in massive practice from the onset of walking, walking up to 8 football fields per day. Second, walking is initially as efficient as crawling. Although newly walking infants often fall, they also travel more distance. This observation changes the framing of questions about motivation: if walking is as efficient as crawling, why not walk? Walking creates no additional cost and has many other advantages, enabling infants to carry objects from one location to another and providing a continuous view of the world as they move.

The lesson we take from this work on locomotor development is that questions about transitions in development must be framed within the context of the full range of infants' experiences. Thus, if we want to understand what motivates the infant to move from familiarity- to novelty-seeking over development, we must connect exploration in the laboratory to exploration in the real world. One approach to connecting up these worlds is to evaluate infants' familiarity with items outside of the lab and assess how they learn about those same classes of items inside the lab. For example, Quinn et al. (2002) found that infants' raised by female caregivers were capable of remembering individual female faces in the lab. Similarly, Kovack-Lesh et al. (2008) found that infants raised with pets in the home were capable of remembering individual cat exemplars in the lab. These findings show empirically that the massive visual experience infants acquire outside of the lab is, in fact, a key driver of development. But these are examples of how infants' experience with specific classes of items outside of the lab influences how they form memories for those same classes of items in the lab. Do massive quantities of visual experience in the real world also impact the more general ability to seek novelty?

We examine this possibility in the present report using a novel approach to understanding visual cognition in infancy—computational modeling. Our starting point is an autonomous Dynamic Neural Field (DNF) model of infant looking and learning developed by Perone and colleagues (Perone et al., 2011; Perone and Spencer, 2013a,b). We have used this model in the past to capture data from studies on the familiarity-to-novelty shift. To do this, we changed parameters of the model over development "by hand" to gain an understanding of how this transition might emerge over development. The insight from this work was that general parameter changes in the strength with which excitatory and inhibitory neurons interact in the model transformed an initially familiarity-seeking model into a novelty-seeking model. The key mechanism underlying this change was the emergence of a new ability—the ability to form a working memory (WM) for

objects. The ability of the model to quickly form working memories for objects enabled it to recognize those objects as known and explore new objects.

Here, we ask if this model can develop itself and show the autonomous emergence of novelty-seeking behavior. In particular, can we initialize a model with a given set of parameters, situate this model in a virtual world, and let it create its own developmental shift from familiarity- to novelty-seeking via autonomous visual exploration. If so, we can then take a step back and ask: what motivated the model to seek novelty?

In the sections that follow, we describe the DNF model and the hypothesis that guided our “by hand” exploration of development in previous work. We then pursue a demonstration proof that the model can develop autonomously through a variant of Hebbian learning. We do this first at a group level. We created a term infant model, let it develop “outside” the lab, and repeatedly brought the model “into the lab” to assess whether it exhibited the familiarity-to-novelty shift in the processing speed task developed by Rose et al. (2002). Results show that the model effectively captures many aspects of the developmental shift. We also asked whether changes in the initial conditions of the model could mimic the development of preterm infants. Results show that the model captures the developmental delays this infant population exhibits.

These simulations provide an initial demonstration that novelty-seeking can emerge from the accumulation of massive out-of-lab experience in our computational model. But intrinsic motivation is not a group-level phenomenon. The motivation to push boundaries in development happens at the level of the individual infant. Thus, in a second study, we looked at the characteristics of individual simulations and ask whether each simulation creates its own unique path from familiarity- to novelty-seeking. These simulation data provide new insights into the sources of individual differences. We conclude by returning to the issue of intrinsic motivation and raise the possibility that no explicit motivational force is needed to explain developmental change within an autonomously behaving complex neural system.

A DYNAMIC NEURAL FIELD MODEL OF INFANT VISUAL EXPLORATION

Figure 1 shows the DNF model architecture. Model equations and parameter values are given in the Appendix. For illustration, the model is situated in a virtual world that consists of a typical laboratory setting in which relevant stimuli appear at left and right locations, task-irrelevant stimuli appear at away locations, and attention-getting stimuli often used to orient infants to the location at which stimuli appear at a center location. The fixation system consists of a collection of nodes that fixate left (L), right (R), center (C), and away (A) locations in a winner-take-all fashion. When a node is suprathreshold (>0), it is said to be in the fixation state. The presence of objects in space bias the fixation system to enter the fixation state (see green arrow from space to fixation system).

The fixation system is reciprocally coupled to a neurocognitive system shown in the bottom panels of **Figure 1**. One component of the neurocognitive system is a perceptual field (PF) that consists of a population of neurons with receptive fields tuned to a

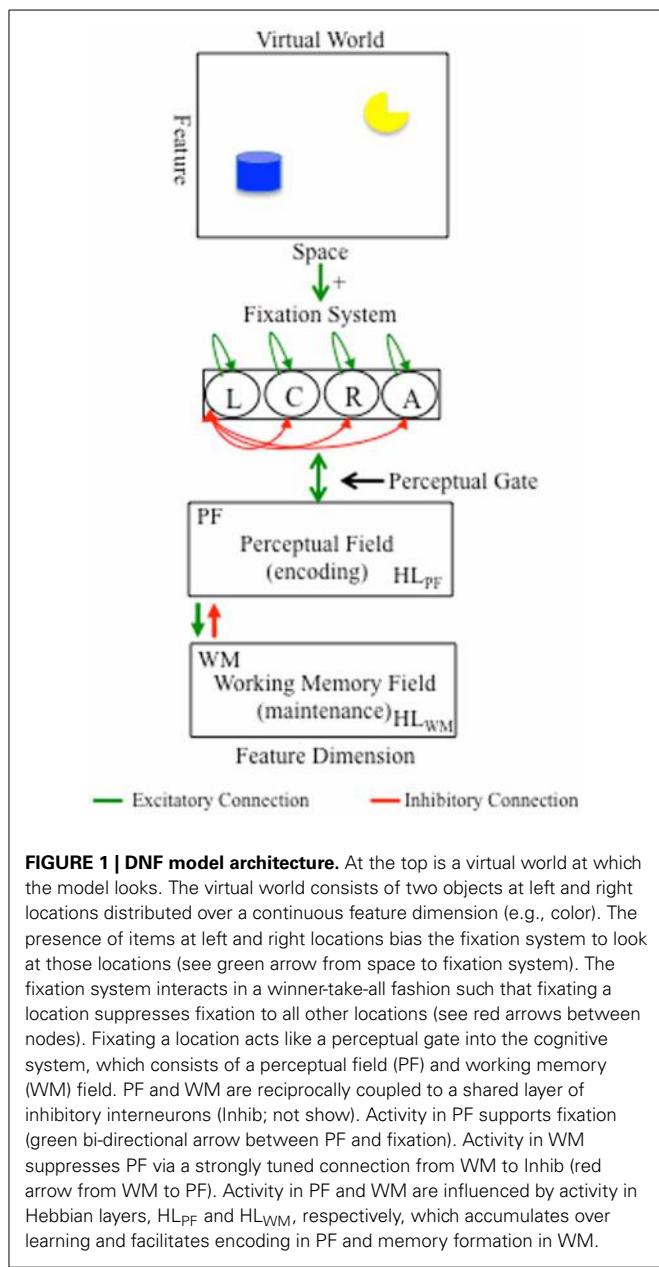


FIGURE 1 | DNF model architecture. At the top is a virtual world at which the model looks. The virtual world consists of two objects at left and right locations distributed over a continuous feature dimension (e.g., color). The presence of items at left and right locations bias the fixation system to look at those locations (see green arrow from space to fixation system). The fixation system interacts in a winner-take-all fashion such that fixating a location suppresses fixation to all other locations (see red arrows between nodes). Fixating a location acts like a perceptual gate into the cognitive system, which consists of a perceptual field (PF) and working memory (WM) field. PF and WM are reciprocally coupled to a shared layer of inhibitory interneurons (Inhib; not show). Activity in PF supports fixation (green bi-directional arrow between PF and fixation). Activity in WM suppresses PF via a strongly tuned connection from WM to Inhib (red arrow from WM to PF). Activity in PF and WM are influenced by activity in Hebbian layers, HL_{PF} and HL_{WM} , respectively, which accumulates over learning and facilitates encoding in PF and memory formation in WM.

continuous feature dimension (e.g., color). The model can represent stimuli along multiple dimensions (see Perone and Spencer, 2013b). For simplicity, we use one dimension here. When a given node in the fixation system is in the fixation state, the stimulus at the associated location is input into PF which encodes the stimulus by forming an activation peak that estimates the feature value (e.g., blue). Neuronal activity within PF is governed by local excitatory/lateral inhibitory interactions. These interactions within PF are relatively weak; thus, once a stimulus is removed, the activation peak relaxes back to the neuronal resting level.

Encoding within PF has two important functions in the model. First, encoding supports continued fixation. Activation in PF feeds back into the fixation system which sustains the fixation state and supports further encoding of the stimulus.

Second, encoding leads to the formation of working memories. In particular, activation in PF passes excitatory input to a layer of similarly tuned neurons in a WM field. Like PF, neuronal activity within WM is governed by local excitatory/lateral inhibitory interactions. Unlike PF, however, neural interactions within WM are stronger. Consequently, activation peaks can be maintained in the absence of input via recurrent excitatory and inhibitory interactions. This is the mechanism for maintaining information in WM in the model.

There are two other patterns of connectivity in the DNF model. First, PF and WM are reciprocally coupled to a shared layer of inhibitory interneurons (Inhib; not shown). This connectivity creates the lateral inhibitory interactions within PF and WM. Critically, the connection from WM to Inhib is set such that strong activity in WM suppresses activity in similarly tuned neurons in PF (see red arrow from WM to PF). This weakens support for fixation from PF, leading to the release from the fixation state when a WM peak is present. Thus, the model encodes a stimulus which drives sustained looking and forms a WM for the stimulus which drives looking away. Second, PF and WM are reciprocally coupled to Hebbian layers (HL; not shown) that implement a form of Hebbian learning. In particular, suprathreshold activity in PF and WM leads to the accumulation of activation at similarly tuned sites in HL_{PF} and HL_{WM} , respectively. The absence of suprathreshold activity in PF and WM leads to slow decay in these HL. Activation traces in HL_{PF} facilitate encoding of previously encoded stimuli in PF. This supports familiarity-seeking and is the basis of recognizing what is known early in development (Wetherford and Cohen, 1973; Fantz, 1974; Perone and Spencer, 2013a). Activation traces in HL_{WM} facilitate the formation of WM peaks. This can lead to the fast suppression of peaks in PF, freeing the model to look away from familiar or known items toward novel items. Thus, this supports novelty seeking.

Figure 2 illustrates the real-time process by which the DNF model learns as it explores objects in a virtual world over time. The top panels show a model that has accumulated little developmental experience exploring items distributed over a color dimension (**A–F**). The bottom panels show the same model after it has acquired more experience (**G–L**). Each panel has the same format. At the top is a collection of objects that the model is exploring over time. The cartoon infant head shows what object is being fixated during each time slice. The next two figures show activation in PF and WM (see black lines and left y-axis) and the strength of experience accumulated in HL_{PF} and HL_{WM} (see red lines and right y-axis).

In **Figure 2A**, the model first looks at the blue object. This excites neurons in PF which, in turn, supports continued fixation and leads to excitation of similarly tuned neurons in WM. The fixation system is stochastic which enables it to spontaneously disengage fixation and shift gaze direction. In **2B**, the fixation system has switched gaze and is now looking at, encoding, and forming a WM for the yellow object. Notice that activity associated with the blue object has subsided within PF and WM; the model is not encoding the blue object or maintaining a WM of the object. In **2C**, the model has again switched gaze and is fixating the blue object and maintains fixation across **2C,D**. This continued fixation enables the model to form a robust peak in WM and acquire a

long-term memory via the HL (see red “bump” of activity in HL_{PF} and HL_{WM} in **2D**). WM activity is also beginning to suppress PF activity to below threshold levels which leads to less support for fixation. Consequently, the model switches gaze. In **2E,F**, the model switches gaze and is fixating, encoding, and forming a WM for the orange object. Once again, the WM of the blue object is not maintained.

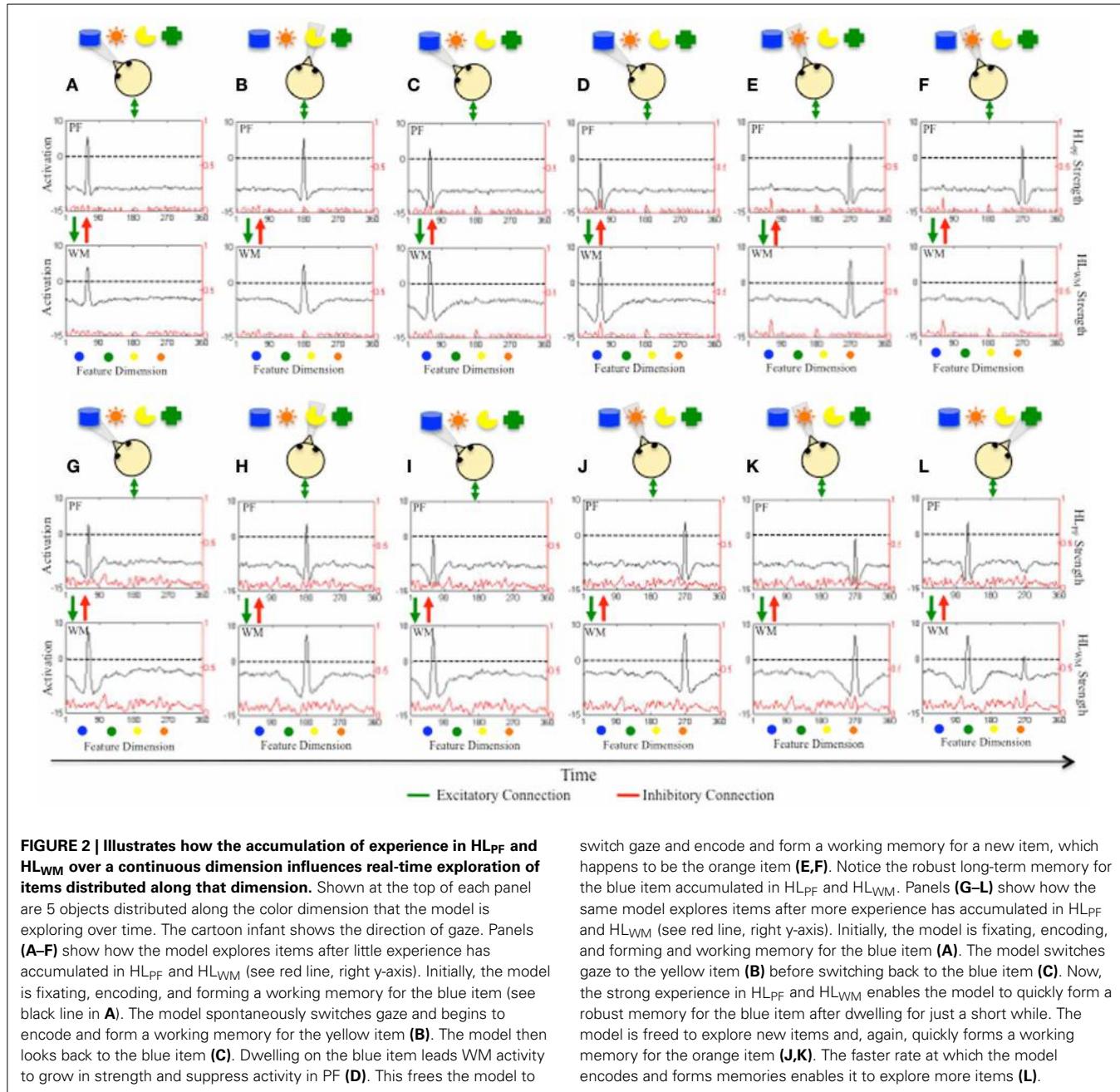
In **Figures 2G–L**, the same model has acquired more experience by exploring a virtual world consisting of objects distributed over a color dimension. This experience has created the stronger, densely distributed traces in HL_{PF} and HL_{WM} shown in **2G–L**. This model is now more familiar with the color dimension. This familiarity has a dramatic impact on looking and learning. In **2G**, the model quickly encodes the blue object into WM, suppressing PF activity to near threshold levels, and biasing the model to switch gaze. In **2H**, the model is fixating the yellow object and, again, WM activity suppresses PF activity to near threshold levels. When the model re-fixates the blue object, WM activity suppresses PF activity to below threshold levels (**2I**) and the model quickly looks away—the model is seeking novelty.

Critically, this novelty seeking behavior is a result of the accumulated long-term experience—the model quickly forms robust working memories because the Hebbian traces have moved WM closer to threshold. This can be seen in **2J–L**. The model fixates the orange object (**J**) and forms a robust memory after maintaining fixation (**K**). This enables the model to explore a new location at which the green object is present (**L**). Notice that WM activity associated with the orange object is hovering around threshold in **2L** even though the model is fixating the green object. This ability to form an enduring, actively maintained WM enables the model to seek novelty, actively contrasting what is known with what is novel. This emerges from a confluence of factors including the duration with which the model fixates an object, the strength of HL_{WM} that facilitates activity within WM, and the strong tuning of local excitatory/lateral inhibitory interactions within WM. This stable WM peak has a dramatic impact of the model’s behavior. For example, when the model re-fixates items that it is actively maintaining in a WM state, PF activity is quickly suppressed. This leads to the quick release of fixation and frees the fixation system to seek novel items.

SIMULATION EXPERIMENT 1

The goal of Simulation Experiment 1 was to probe whether the model could develop novelty-seeking behavior from autonomous visual exploration in a “real” world. If so, this might shed light on where the motivation to seek novelty comes from. As described previously, this goal emerged from our earlier work using the DNF model to quantitatively simulate the familiarity-to-novelty shift in early development (Perone and Spencer, 2013a). We did this by changing parameters of the model “by hand” according to the spatial precision hypothesis (SPH) proposed by Schutte and Spencer (2009; see also Schutte et al., 2003; Simmering et al., 2007; Perone et al., 2011; Perone and Spencer, 2013a,b).

According to the SPH, excitatory and inhibitory interactions become stronger over development, leading to more robust neural activation states and “sharper” peaks of activation. Implementing the SPH involves strengthening within-layer



excitatory connections in PF and WM and cross-layer inhibitory interactions from Inhib to PF and WM. When neural interactions are weak, the model slowly encodes and slowly forms peaks in WM. This creates a familiarity-seeking model that dwells on familiar items for relatively long durations before looking to novel items. When neural interactions are stronger, the model quickly encodes items and quickly forms peaks in WM. This creates a novelty-seeking model that has short dwell times on familiar items before looking to novel items (see Perone and Spencer, 2013a).

Implementing the SPH in the DNF model only requires changes in the strength of excitation and inhibition. Might these changes emerge from a simple Hebbian learning process? Recall

that HL coupled to PF and WM accumulate memory traces as peaks are built in PF and WM. This increases the excitability of previously active sites as well as nearby sites based on a similarity gradient. As general experience across a dimension accumulates, this might approximate the increase in excitatory strengths we implemented by hand. What about the increase in inhibition? As excitatory interactions strengthen, PF and WM will pass stronger activation to the shared inhibitory layer. This might give rise to an effective increase in inhibition as well.

We explore these possibilities here across two groups of simulations. In one set of simulations, the DNF model was tuned to mimic the behavior of term infants. In the second set of simulations, the model was tuned to be “less mature” using the SPH as

a guide. This enabled us to examine how the initial conditions set by the model parameters impact development relative to the role of massive “out-of-lab” experience. To benchmark these simulations, we assessed the familiarity- and novelty-seeking biases of the model in the processing speed task developed by Rose et al. (2002) by repeatedly bringing the model “into the lab” over the course of its development.

Figure 3 shows a schematic of the processing speed task. At the beginning of the task, infants are presented with a pair of different stimuli. In Rose et al. (2002), faces were used as stimuli. The procedure has been used in other studies as well and is robust to variation in stimuli (Robinson and Sloutsky, 2007). After the first trial, one item was designated as the familiar item and remained unchanged across trials (orange star). Infants were required to accumulate 4 s of looking on each trial. Once infants met the looking criterion, the trial ended and the next trial began. On each trial, a novelty score was calculated by dividing looking to the novel stimulus by total looking accumulated across the novel and familiar stimulus. The measure of processing speed was the number of trials required to exhibit a novelty score greater than 55% on three consecutive trials.

Rose et al. (2002) reported three additional measures of looking. The first is looking to the familiar item which is the amount of time infants accumulated looking to the familiar stimulus prior to meeting criteria. This is a good index of infants’ familiarity seeking bias and has long been assumed to reflect the time

required for infants to form memories (Cohen, 1972a,b; Hunter and Ames, 1988; Colombo and Mitchell, 1990). The second is shift rate which is the rate of gaze switching relative to time spent looking. Shift rate has been proposed to reflect the efficiency with which infants distribute their attention through time and space (Rose et al., 2007). The last is look duration which is the average length of each look. Like shift rate, look duration has been proposed to be a measure of disengaging and distributing attention (Rose et al., 2007).

Figures 4A–D shows infants’ performance in the processing speed task (Rose et al., 2002). The left portion of each panel shows term infants at 5 months of age (blue bars), 7 months of age (red bars), and 12 months of age (black bars). Over development, term infants exhibited a decrease in trials to reach criterion (A), accumulated less time looking to the familiar item (B), exhibited higher shift rates (C), and exhibited shorter look durations (D). Preterm infants produced a similar pattern of results but, critically, at each age exhibited behavior that resembled relatively younger term infants. For example, 7-month-old preterm infants required about the same number of trials to reach criterion as 5-month-old term infants. This pattern of results indicates that preterm infants are delayed on these measures.

In the past, we have used the DNF model and SPH to capture developmental changes in the suite of measures assessed by Rose et al. (2002) using data from a preferential looking paradigm (Perone and Spencer, 2013a). Here, we test whether the accumulation of experience in the DNF model can do the work of the SPH and quantitatively simulate the empirical data shown in **Figure 4**.

METHOD

The DNF model was situated in a simple virtual world consisting of two items that varied along a single dimension. The dimension consisted of 360 degrees of metrically organized continuous feature space (e.g., color). We randomly sampled items for the model to explore from a set of 360. A non-fixated item was replaced every 1000 time steps. This enabled the model to sample many different items over time, consistent with what infants might experience interacting with parents as they show infants different toys from a larger set of possible toys.

The simulations were parsed into 30 10,000 time step episodes of visual exploration (300,000 time steps of experience in total). Conceptually, these episodes occur over the time scale of months; however, in the model, we condensed this experience considerably to keep the simulation time reasonable (e.g., even with this condensed “out-of-lab” experience, it took over 8 h of simulation time to run a single simulation through the full set of out-of-lab and in-the-lab experiences). In addition to the 30 episodes of exploration, we inserted inter-episode intervals of 100 time steps. During these intervals PF, WM, and Inhib were re-initialized (i.e., set to 0 activation). This eliminated any sustained WM peaks and reset the fields for exploration of new items at the onset of the next episode.

We wanted to test whether differences in the initial conditions of the DNF model could account for population differences in the familiarity-to-novelty shift over development. Thus, we created two models with differences in the initial parameter values using the SPH as our guide. Specifically, we first created a “term”

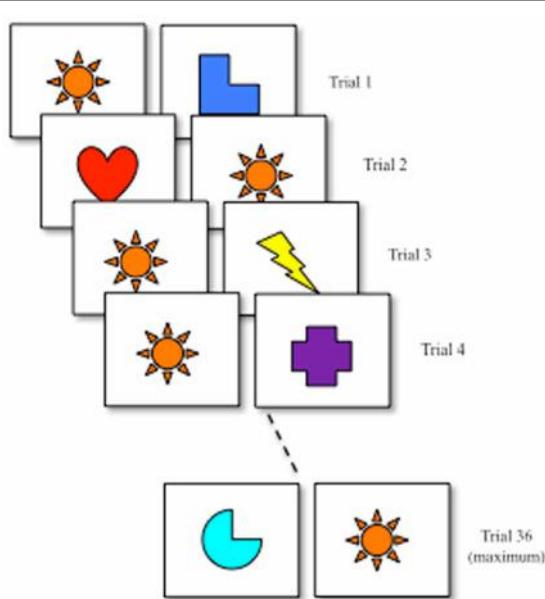


FIGURE 3 | Processing speed task developed by Rose et al. (2002). Infants were presented with a pair of different stimuli on each trial. Across trials, one stimulus remained unchanged (familiar) and one changed (novel). On each trial, infants were required to accumulate 4 s of looking. Infants met a learning criterion once they looked at the novel stimulus more than 55% of the time on the 3 consecutive trials or 36 trials had passed. In the empirical study, stimuli were faces. There were 19 stimuli, one designated as the familiar and 18 designated as novel. If 18 trials had passed before infants met the criteria, the 18 novel stimuli were represented.

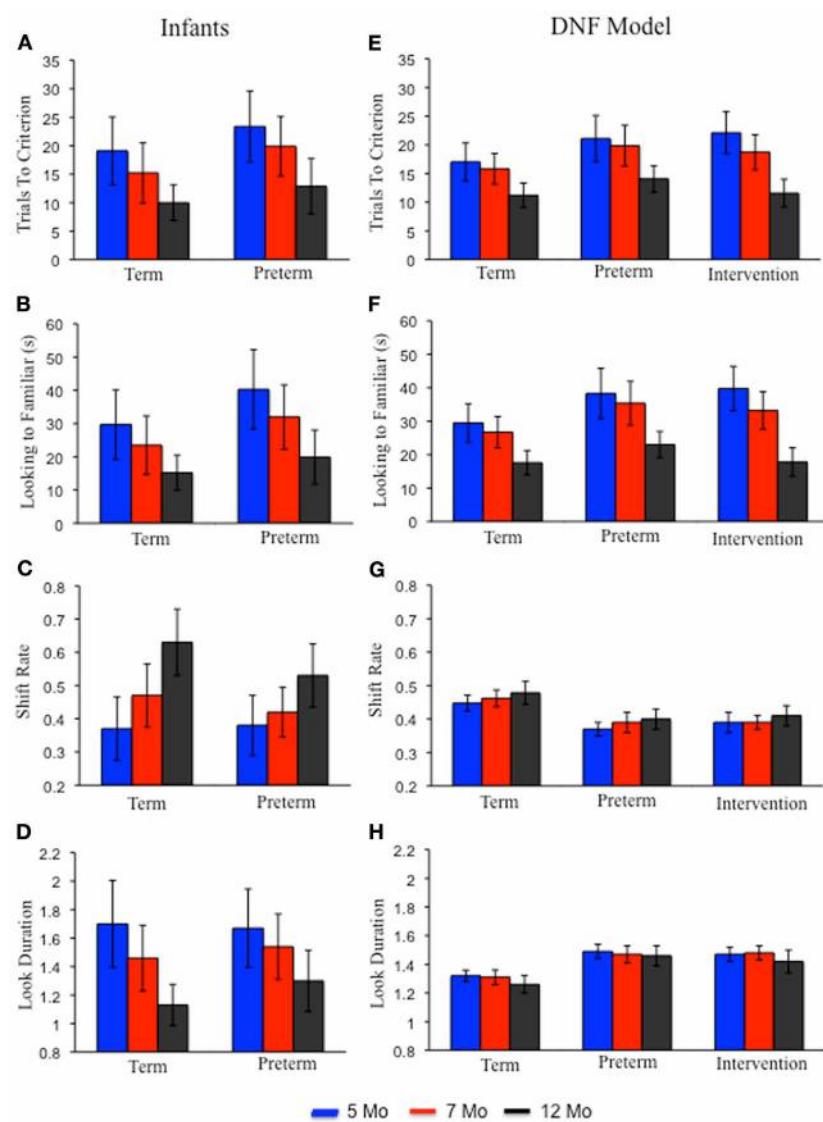


FIGURE 4 | Panels (A–D) show empirical results from the processing speed task reported by Rose et al. (2002) for term (left) and preterm (right) infants at 5 (blue), 7 (red), and 12 (black) months of age. With age, term and preterm infants exhibited fewer trials to criterion (A), accumulated less time looking to the familiar (B), higher shift rates (C), and shorter look durations (D). Preterm infants' behavior at every age

resembled that of younger, term infants. Panels (E–H) show results from the DNF model in the processing speed task for the term (left), preterm (middle), and intervention (right) models. The DNF model exhibited a similar pattern of results for the term and preterm infant models. The intervention model showed performance that resembled the term model by 12 months of age.

model. To do this, we allowed the DNF model to accumulate experience in the HL by exploring a virtual world and assessed its performance in the processing speed task over the course of its development (see below). We then hand-tuned the DNF model parameters until we established a parameter set that produced a pattern of results that was quantitatively similar to the empirical results for the term infants reported by Rose et al. (2002). After that, we uniformly weakened the SPH parameters by 20%. We will refer to this weaker parameter set as the “preterm” model.

The development of the term and preterm models were simulated 5 times each. During each simulation, we saved the state of HL_{PF} and HL_{WM} after each episode of exploration. We then

averaged HL_{PF} and HL_{WM} across all 5 simulations. This created nearly uniform levels of activation across all neuronal sites in the HL by smoothing out the peaks and valleys of activation in the layers that were unique to each individual simulation (e.g., compare the HL for group level simulations in Figure 5 to individual simulations in Figure 7). This uniformity mimics the strengthening of excitatory connections across an entire dimension we implemented by hand when we implemented the SPH in previous work. Our goal in averaging the HL was to maximize the stability of the model’s behavior across simulations when situated in the processing speed task (see below), much like averaging the looking behavior across a group of infants.

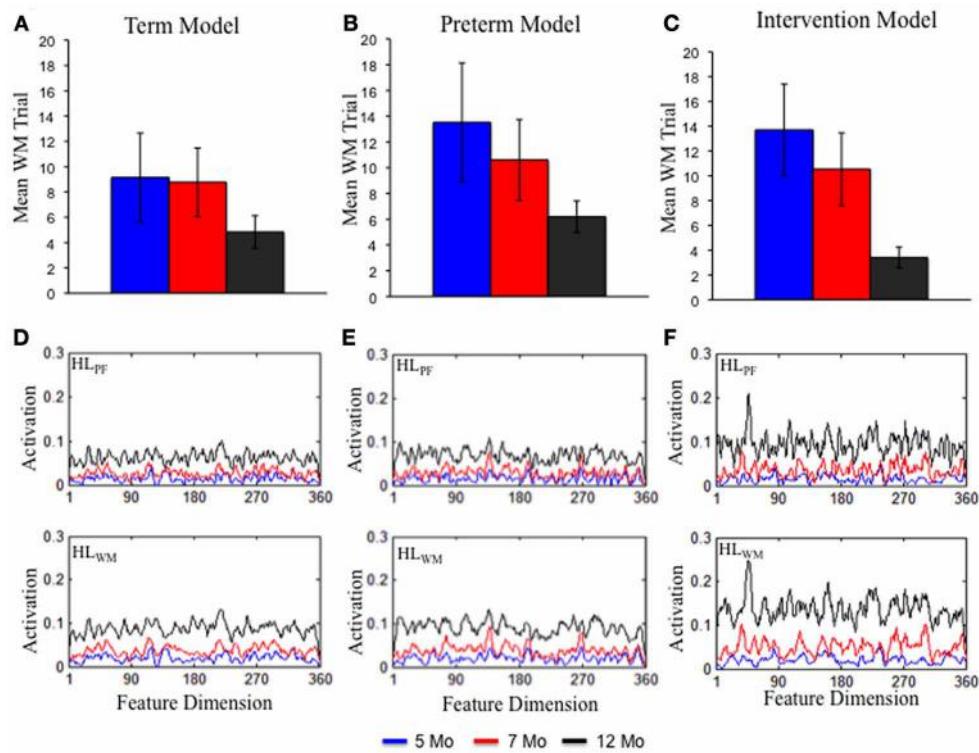


FIGURE 5 | The top shows the rate at which DNF model formed a stable WM peak for the term (A), preterm (B), and intervention (C) models at 5 (blue), 7 (red), and 12 (black) months of age. Over development, all models formed a stable WM peak more quickly. The rate of WM peak formation was delayed for the preterm model but

enhanced by 12 months of age for the intervention model. The bottom shows the experience accumulated in HL_{PF} and HL_{WM} for the term (D), preterm (E), and intervention (F) models. The strength of activation in the Hebbian layers was comparable for the term and preterm infant models. It was stronger for the intervention model.

Next, we initialized the term and preterm models with their respective mean HL_{PF} and HL_{WM} accumulated at 5, 10, and 30 episodes and situated each model in the processing speed task developed by Rose et al. (2002). For ease of comparison to the empirical data, we refer to these initializations as the term and preterm infant models at 5, 7, and 12 months of age. We ran 100 simulations of each model. This number of simulations provided a thorough assessment of the range of the model's looking behavior in the processing speed task in the context of the natural variation the model shows when placed in a laboratory-based learning task (for a discussion, see Perone and Spencer, 2013a,b). To precisely map the models' performance in the lab with the timing of events in the speed of processing task, we assumed the mapping used in our previous studies where 200 time steps in the model was equal to 1 s (Perone and Spencer, 2013b). Note that in the simulation method described above, learning inside the lab did not influence the model's performance outside of the lab.

RESULTS AND DISCUSSION

The simulation results are presented in the following three sections. In the first section, we describe the DNF model's performance in the processing speed task and the underlying neurocognitive dynamics. In the second section, we probe whether the development of the preterm infant model might be modified through an intervention. This helped us assess the

influence of the initial model parameters relative to the accumulation of out-of-the-lab experiences. In the third section, we probe whether the accumulation of experience in the HL led to sharper and more robust WM peaks consistent with the SPH we implemented "by hand" in previous work.

Cognitive and behavioral dynamics

Figures 4E–H shows the DNF model's performance in the processing speed task. Like infants, over development the term infant model exhibited a decline in trials to reach criterion (E) and accumulated less time looking to the familiar item (F). The model also showed a small, quantitative increase in shift rate (G) and decrease in look duration (H) over development. Like infants, the preterm infant model exhibited a similar, but delayed, pattern relative to the term infant model.

What are the sources of these developmental changes in the model's performance? The top portion of Figure 5 shows one critical change—the mean trial on which the model first formed a stable WM peak for the familiar item. A stable WM peak was defined as sustaining suprathreshold activity across the inter-stimulus interval (4 s; see Perone and Spencer, 2013a). Over development, the term (A) and preterm (B) infant models form WM peaks more quickly, with the preterm model lagging the term model. This index of the model's performance is important because maintaining the familiar item in WM produces strong

inhibition in PF at sites involved in encoding the item. This, in turn, leads to less looking to the familiar item and more looking to the novel item. That is, quick WM formation allows the model to actively recognize what is known and seek novelty. In addition, quick WM formation leads to more frequent gaze switching and shorter look durations over development, allowing the model to more effectively explore items in the task space.

What drives these changes in WM in the model? These developmental differences emerge from the accumulation of distributed memory traces in HL_{PF} and HL_{WM} over time. **Figures 5C,D** shows the state of HL_{PF} and HL_{WM} for the 5-(blue lines), 7- (red lines), and 12-month-old (black lines) models. Over development, activation across the dimension grew in strength for the term (**C**) and preterm (**D**) models. In other words, the model became increasingly familiar with the entire dimension. This, in turn, led PF to encode items more quickly and WM to maintain those items more robustly.

These simulations shed new light on the origins of intrinsic motivation. Specifically, the simulations allow us to ask where the motivation to seek novelty comes from. Novelty seeking has some distinct advantages over familiarity seeking for infants. For example, novelty seeking enables infants to compare known with unknown items, efficiently explore complex environments, and, more generally, opens the door to discovery. Critically, infants do not know this ahead of time. Our simulations indicate that the motivation to seek novelty emerges from the accumulation of visual experience within a complex, dynamical exploratory system. A key property of the DNF model is that real-time, autonomous exploratory behavior creates a history that influences the behavior of the system at future points in time. The accumulation of this history over time led to the emergence of a new ability—quickly forming working memories of “known” items. This cognitive ability enables an increasing bias to seek novelty to gradually emerge without an explicit motivational force. We discuss this topic further in the General Discussion.

These simulations also shed new light on the population differences in the familiarity-to-novelty shift. In particular, the Hebbian traces accumulated for the term and preterm model were quite similar (compare **Figures 5D,E**) and were not sufficient to overcome the weaker neural interactions in the preterm infant model. This indicates that population differences in visual exploration and WM formation are largely attributable to the initial conditions of the system, while developmental changes emerge from the accumulation of out-of-the-lab experiences. Below, we probe whether altering the experience of the preterm infant model during development influences its novelty seeking behavior in the processing speed task.

Intervention

The simulations results described above show that novelty seeking emerges as experience accumulates via a Hebbian learning process. However, the initial conditions of the model played a major role in development: the accumulation of experience did not enable the preterm model to overcome the initially weaker neural interactions. How strong is this constraint on development? Are there ways that we might enhance the model’s experience and, in turn, foster the development of novelty seeking biases?

There is a large literature showing that how other agents (e.g., parents) interact with infants while exploring objects influences how they distribute their looks (Landry and Chabieskie, 1988; Perrinello and Ruff, 1988). This is especially salient in interventions with preterm infants. For example, Landry et al. (2006, 2008) have shown that preterm infants benefit in the social, cognitive, and linguistic domains when parents are trained to act responsively to their infants while exploring objects as part of an intervention. This involves “following in” on the objects infants explore and helping infants maintain attention (e.g., by manipulating the object of infants’ focus) rather than shifting attention to other objects (e.g., manipulating an object elsewhere).

Can we manipulate the nature of the preterm model’s experience and transform it into a term-looking model in a similar way? For example, can we bias the model to continue looking at an object and, in turn, enhance encoding, WM, and long-term memory formation? Could this enhance traces in HL_{PF} and HL_{WM} enough to overcome the weaker neural interactions of the preterm model? This would help us assess the relative impact of the model’s initial parameter setting versus the accumulation of out-of-lab experiences.

To test this possibility, we re-simulated the development of the preterm model. After the fifth episode, we implemented an intervention. We wanted to probe how an intervention might unfold in the real world where infants do some developing during the first few months of life, undergo assessment, and are assigned to an intervention thereafter. In our intervention, the model was biased to sustain looking at whatever item it happened to be fixating. If the model was fixating the left location, for example, the input from the object in at that location in space was increased. This, in turn, provided a slight boost of excitation to the fixation system, helping to maintain fixation. In the DNF model, this is equivalent to another agent manipulating an object in space (see **Figure 1**).

Figure 4 shows the simulation results. The intervention had the most dramatic impact on the number of trials to criterion (**E**) and looking to the familiar item (**F**). In particular, by 12 months the intervention model met criterion at a rate comparable to the term model at the same age. Similarly, by 12 months the intervention accumulated less time looking to the familiar item much like the term model at the same age. A substantive amount of intervention experience was required for the intervention to exert its effects on the model’s performance in the processing speed task. Ultimately, however, the intervention created a preterm infant model with a robust novelty-seeking bias comparable to term infants.

What are the sources of these behavioral changes? **Figure 5C** shows the trial on which the intervention model formed a stable WM peak. At 5 (blue bars) and 7 (red bars) months, the intervention model formed a WM peak at rates comparable to the preterm infant model (**B**). By 12 months (black bars), however, the intervention model formed a WM peak at rates that exceeded the term model (**A**). This improved capacity of the intervention model to quickly encode items and maintain items in WM arises from the strength of activity accumulated in the HL. As can be seen **5D**, the strength of HL_{PF} and HL_{WM} by 12 months (black lines) is much stronger than at the same time for the term (**C**) and preterm (**D**) infant models. This stronger accumulation of activity in the HL

enabled the intervention model to overcome the weaker neural interactions of the preterm infant model.

Spatial precision hypothesis

In our previous work, we implemented the SPH by hand, showing that stronger neural interactions lead the DNF model to form working memories more quickly (Perone and Spencer, 2013a). This effective increase in processing speed also led to stronger biases for novelty, shorter looks, and higher rates of gaze shifting. Here, we observed these very same patterns of change over development. But does the accumulation of experience via Hebbian learning yield the same changes in neural interactions produced by SPH?

Implementing the SPH via hand-tuning neural interactions leads to stronger, narrower WM peaks with deep lateral inhibition (see Schutte and Spencer, 2009). We tested whether the accumulation of experience in the HL reproduced this activation profile by initializing the DNF model with the state of the HL after 5, 10, 15, 20, 25, and 30 episodes of exploration. The model was presented with a single stimulus for 2000 time steps. When the stimulus was removed, we sampled the state of WM every time step for 1000 time steps. We then averaged the state of WM across all samples to obtain a representative WM profile. Noise was turned off so that we could obtain a clean estimate of how the HL impact WM peaks (see Schutte and Spencer, 2009).

The results are shown in **Figure 6**. Over development, the strength of the activation peak increased. After 5 (red), 10 (blue), and 15 (green) episodes of exploration, the peak was too weak to maintain a stable WM peak under the task conditions. After 20 episodes of exploration (cyan), the accumulated memory traces in HL_{WM} enabled WM to maintain a peak at suprathreshold (>0) levels. The model effectively acquired a new cognitive

ability. In addition, the excitatory component of the peak grew in strength and became somewhat narrower over development. The inhibitory component grew broader and deeper as well. It is notable that these neurodevelopmental changes in excitation and inhibition were all driven by the accumulation of excitatory memory traces. As the strength of HL_{WM} increased, excitation in WM became stronger which passed stronger activation into the layer of inhibitory interneurons. This, in turn, projected stronger lateral inhibition back to WM. Thus, the present simulations demonstrate that the SPH can emerge over development via a variant of Hebbian learning as the model accumulates “out-of-the lab” experiences.

Simulation Experiment 1 revealed three key insights. First, the accumulation of visual experience along a dimension leads to quicker WM formation for stimuli on a familiar dimension. This quick recognition, in turn, promotes novelty-seeking. Second, the impact of visual experience on cognition is influenced by the initial state of the neurocognitive system. The neurocognitive deficits of the preterm infant model were expressed over development, leading to slower WM formation along a familiar dimension across the first year. Increasing the intensity of the experience the preterm infant model acquired with a dimension, however, enhanced WM formation by strengthening the familiarity with that dimension. Lastly, the accumulation of visual experience led to stronger neural interactions within the neural populations involved in encoding and WM formation. This strengthening was created by the accumulation of Hebbian learning but resembled the SPH at the neurocognitive (faster WM formation) and behavioral (less looking to familiar items) levels. These results indicate that processing speed and, consequently, the transition to novelty seeking over development emerges from experience.

SIMULATION EXPERIMENT 2

Simulation Experiment 1 showed that the familiarity-to-novelty shift emerges over development as experience accumulates via a Hebbian learning process. It also showed that the motivation to seek novelty comes for free from the dynamics of a historical cognitive and behavioral system. But these simulations were at the level of the group. Recall we simulated the development of 5 individuals and initialized the model in the processing speed task with the average state of those individual HL. The motivation to push boundaries in development, however, happens at the level of the individual. Each individual must forge a unique path and strive beyond what is currently possible.

In the infant cognition literature, individual differences in visual exploration have long been attributed to differences across infants in the neurodevelopmental mechanisms that underlie basic perceptual and cognitive processes (Colombo and Mitchell, 1990; Rose et al., 2007). This position stems from two observations. First, numerous studies have shown that individual differences in looking are stable during the first year of life (Colombo et al., 1987; Rose et al., 2001). Second, individual differences in looking are predictive of cognitive developmental outcomes in toddlerhood (2009) and adolescence (2012).

This view of individual differences is generally consistent with the group-level differences from Simulation Experiment 1. There,

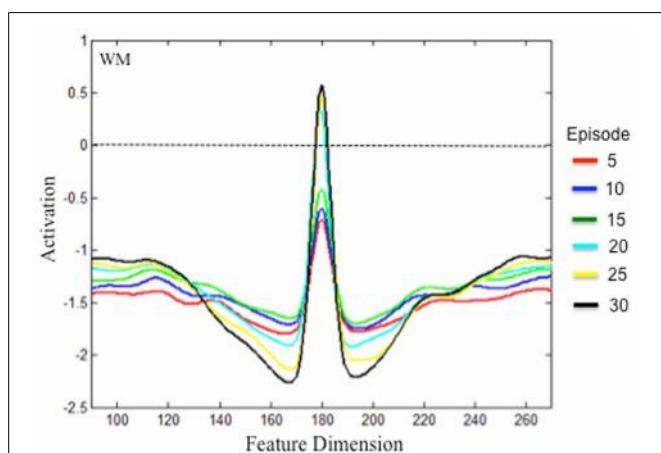


FIGURE 6 | Test results of whether experience accumulated across a dimension can lead to the SPH at the level of neural interaction. The model was initialized with the experience accumulated in the Hebbian layers after every 5 episodes of exploration, which is shown by the different colored lines. The figure shows the state of WM during the inter-stimulus interval following stimulus presentation (see text). With experience, the WM field was able to form a stable peak. This peak had a strong excitatory component and deep inhibitory component much like implementing the SPH via hand-tuning the strength of excitatory and inhibitory connections.

differences across simulations reflected, in part, different initial conditions in parameter values. Applied at the level of individuals, we might create an entire ensemble of individual models, with some models starting off with slightly stronger excitatory and inhibitory interactions than others (see, e.g., Perone and Spencer, 2013a). But individual differences might also reflect the differential accumulation of experience over development. For instance, Perone and Spencer (2013a,b) showed that experience on the task time scale creates variation in looking that mimics aspects of developmental changes, even when models start with the same initial conditions. Might the accumulation of experience over development lead to stable individual differences even when models—or infants—start out in the same neurodevelopmental state? We probe this possibility below.

METHOD

We simulated the development of 10 individuals term, preterm, and intervention models using the same method described above with one exception. In the simulations above, we averaged the HL across simulations and situated the model in the processing speed task after 5, 10, and 30 episodes. Here, we initialized the model with HL_{PF} and HL_{WM} from each of the 10 individual simulations. As above, each model was run in the processing speed task 100 times to assess the full range of performance for each individual.

RESULTS AND DISCUSSION

Figure 7 shows a sample of three individual term infant simulations. The left portion shows the activation traces in HL_{PF} and HL_{WM} for each simulation. Notice that each simulation varies in the strength, distribution and location of peaks and valleys along the feature dimension. Also notice that these peaks and valleys are much more pronounced at the individual level than at the group level (compare 7A to 5C). This highlights individual differences in what the model happened to form robust memories for during its development. The right side of the figure shows three measures from the processing speed task: trial of stable WM peak formation, trials to criterion, and looking to the familiar. As can be seen, each individual follows a distinct, yet similar, developmental trajectory. For example, the individual in 7A showed a shallow, steady decrease in the trials to meet criterion over development. The individual in 7B showed a steep decline. And the individual in 7C showed little decline from 5 to 7 months but a sharp decline from 7 to 12 months.

This holds true for the preterm infant model as well. Three individual simulations of this model are shown in **Figures 8A–C**. Consistent with the group level simulations, the structure of the developmental trajectories for the individual term and preterm infant models were influenced by the initial conditions. That is, individual preterm infants exhibited a similar, yet delayed, developmental trajectory relative to the individual term infant models. The pattern is somewhat different for the intervention model. Three individual simulations of this model are shown in **Figures 9A–C**. For the intervention simulations, some individuals showed a dramatic decline in trials to criterion by 12 months of age, much like the group level analyses (see 9C). Others, by contrast, showed an increase in the number of trials to criterion (see 9A).

Figures 7–9 show that each individual had a unique developmental trajectory. But did the accumulation of experience in the model create a stable pattern of familiarity and novelty seeking biases over development? In other words, were familiarity-seeking individuals early in development also familiarity-seeking individual later in development? **Figure 10** shows the trials to criterion for the 10 individual term, preterm, and intervention simulations. Inspection of the plots reveals some stability over development in each group, even though individual runs of the model in each group had exactly the same initial conditions. For the term infant model, S8 (salmon) and S5 (turquoise) are relatively slow processors at 5 and 7 months. S1 (blue) and S7 (light blue) are fast processors at 5 and 7 months. And S3 is neither fast nor slow at 5 and 7 months. The preterm infant model is considerably more variable. The weaker neural interactions of the preterm model make it more susceptible to stochastic influences. Nevertheless, S3 (green) and S6 (orange) are faster than S9 (light green) and S10 (purple) at all three ages. The intervention model was even more variable than the preterm infant model, yet it also showed signs of stability. For example, S10 (light purple) was faster than S6 (orange) at all three ages.

The striking variability in the individual intervention simulations indicates that the intervention did not impact every individual in the same way. For example, S4 (dark purple) and S10 (light purple) were both quick novelty-seekers by 12 months. By contrast, S1 (blue) quickly met the novelty-seeking criterion at 5 months but exhibited an increase in trials to criterion at 12 months. **Figure 8A** shows the accumulation of activation in the HL for this model. As can be seen, S1 acquired some tall, broad memory traces (see near site 80) between 7 (red line) and 12 (black line) months in both HL. This pattern of activity can lead to the model to dwell because the traces in PF are so strong. Consequently, the model spent more time looking to the familiar item and exhibited longer look durations at 12 months (black bars) than at 7 months (red bars) even though it actually formed a WM peak more quickly at 12 months than at 7 months. The accumulation of activity in the HL for S5 and S10 are shown in **Figures 8B,C**, respectively. These simulations acquired a more evenly distributed pattern of activity, especially in HL_{PF} . This, in turn, led these simulations to exhibit a relatively consistent shift from familiarity to novelty seeking over development that aligned well with their developing capacity to form working memories. These simulation results raise the exciting possibility that we can map individual models to individual infants and capture the impact of real-world interventions. We return to this issue below.

The results from individual simulations suggest that individual experiences can give rise to stable individual differences over development. To quantify this across the full set of simulations, we used hierarchical regression. **Table 1** shows the regression analysis. The table presents the predictor variables entered on each step and a number of summary statistics. On the left are the proportion of variance accounted for by the predictors (R^2), change in proportion of variance accounted for across steps (change in R^2), change in the F statistic across steps, and the probability associated with the F statistic. On the right are the unstandardized beta weights (β) and standardized beta weights (beta). The weight is the unique contribution of each predictor. The sign indicates

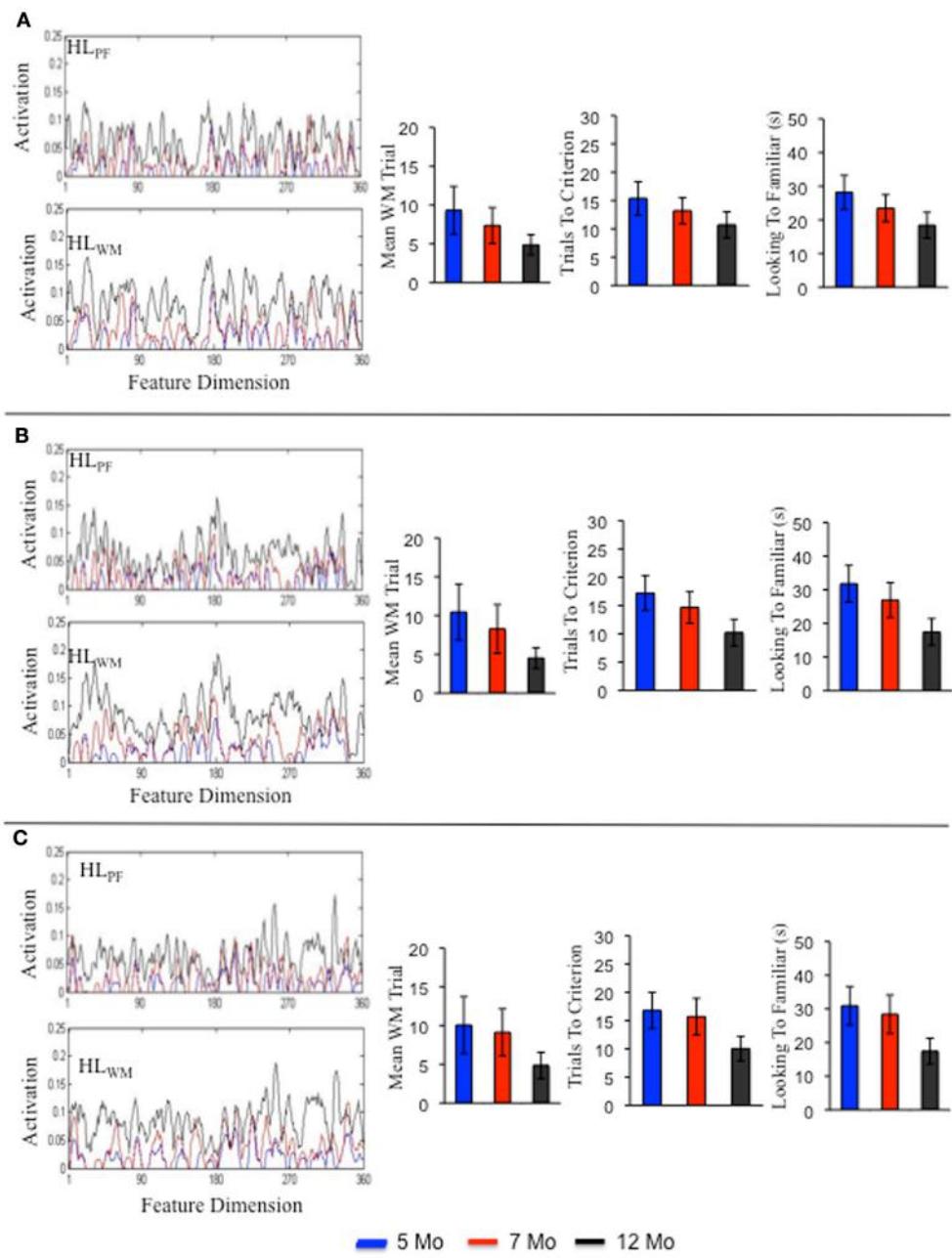


FIGURE 7 | Shows the Hebbian layers and performance in the processing speed task for three individual simulations of the term model. Each panel shows an individual simulation at 5 months (blue), 7 months (red), and 12 months (black).

the direction of the relationship between the predictor and dependent measure. The size of the weight indicates the slope. Steeper slopes indicate that the dependent measure changes more for each unit change in the predictor. The *p* value shows the statistical significance of each predictor.

In the first step, we controlled for group by entering group (term = 1, preterm = 2, and intervention = 3) as a predictor and trials to criterion at 12 months of age as the dependent measure. Group accounted for a significant proportion of variance in trials to criterion at 12 months of age, $R^2 = 0.39$. In the second

step, we entered trials to criterion at 5 and 7 months. Trials to criterion early in development did account for a significant proportion of variance at in trials to criterion later, R^2 Change = 0.19. Evaluating the beta weights indicates that trials to criterion at 7 months of age was the strongest predictor. The positive slope of the beta weight indicates that more trials to criterion at 7 months of age was associated with more trials to criterion at 12 months of age. In the past, we found that experience in the DNF model on the task time scale leads to patterns of covariation between looking and novelty preferences like real infants. These results provide

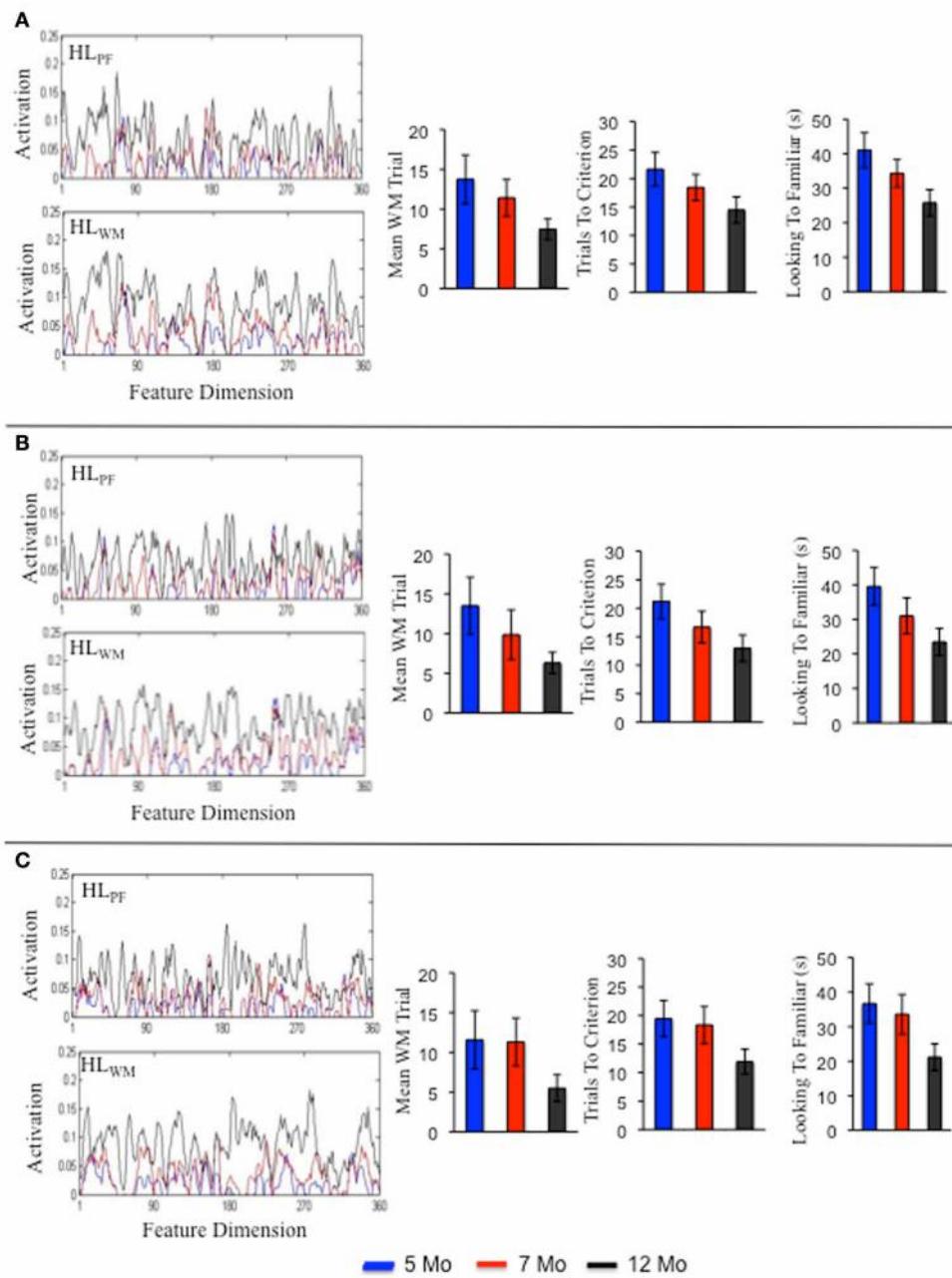


FIGURE 8 | Shows the Hebbian layers and performance in the processing speed task for three individual simulations of the preterm model. Each panel shows an individual simulation at 5 months (blue), 7 months (red), and 12 months (black).

compelling evidence that experience creates stability on the developmental time scale in familiarity and novelty seeking behavior at the level of the individual.

GENERAL DISCUSSION

Children make astonishing transformations during just a short period of time, raising the question of why they continually strive forward in development. Examining the sources of intrinsic motivation early in development might offer a particularly compelling case that provides insights into the very

origins of motivational states. Here, we examined a key transition in exploratory biases in the first year of life as infants move from familiarity-seeking to novelty-seeking. This familiarity-to-novelty shift emerges gradually over the first year, differs across infant populations, and is stable within individuals over time (see Hunter and Ames, 1988; Rose et al., 2001, 2002, 2007). Novelty seeking has some distinct advantages. For example, it allows infants to compare and contrast known items in memory with new items in the environment (Oakes et al., 2008). This might help them form categories and inspect multiple items before

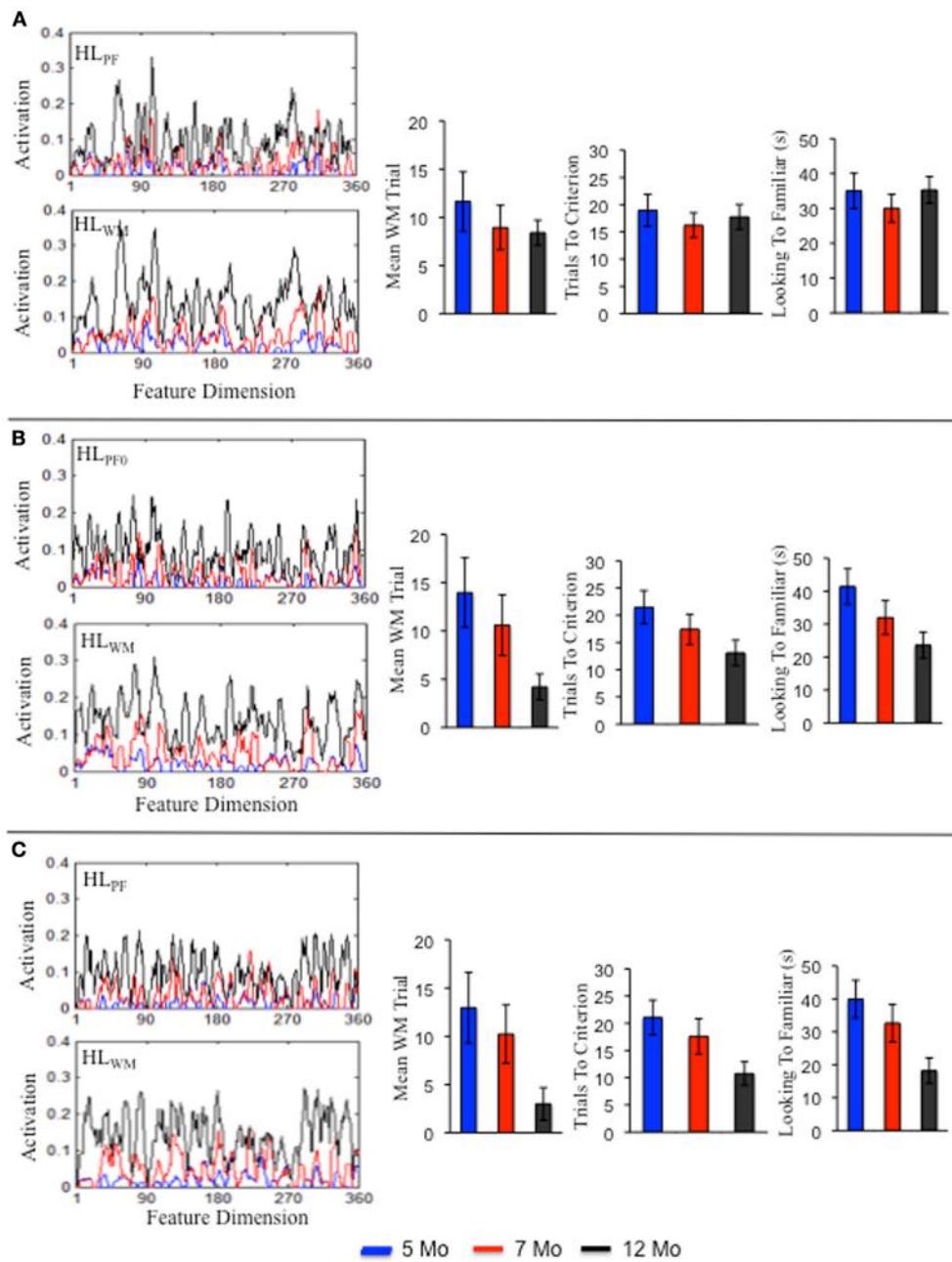


FIGURE 9 | Shows the Hebbian layers and performance in the processing speed task for three individual simulations of the intervention model. Each panel shows an individual simulation at 5 months (blue), 7 months (red), and 12 months (black).

deciding to approach them for further exploration. But what motivates the infant to switch exploratory styles?

To address this question in the present report, we used a DNF model of infant visual exploration that has accounted for the familiarity-to-novelty shift in previous work (Perone and Spencer, 2013a,b). Previous findings showed that when we implemented the SPH “by hand” over development, the DNF model could capture the qualitative and quantitative aspects of this shift. This included examples of infants’ robust familiarity preferences during the first two months of life (Wetherford and

Cohen, 1973; see also Fantz, 1974), as well as the more gradual increase in novelty seeking over the course of the first year. Here, we asked if the DNF model could transform itself from a familiarity to novelty seeking model through nothing more than “out-of-lab” experience. Our strategy was to let the DNF model accumulate experience in HL via autonomously exploring a virtual world consisting of objects distributed over a continuous feature dimension. We then asked whether the model exhibited the familiarity-to-novelty shift in the processing speed task.

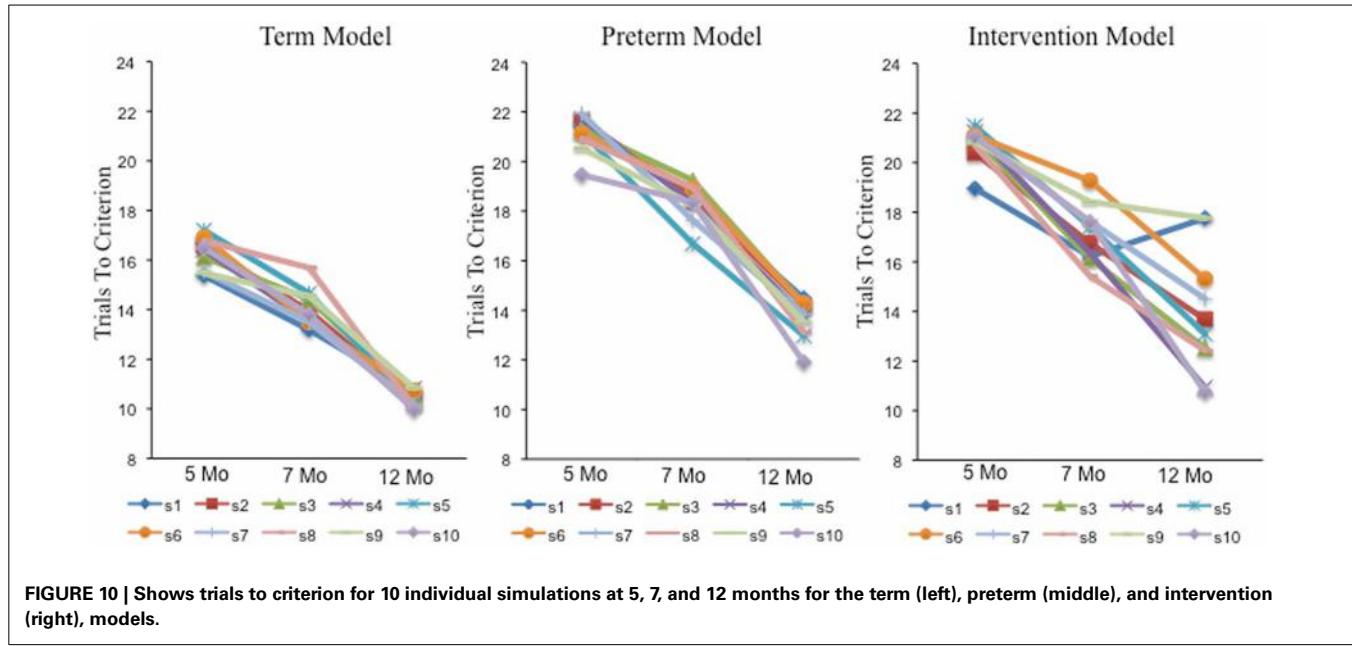


FIGURE 10 | Shows trials to criterion for 10 individual simulations at 5, 7, and 12 months for the term (left), preterm (middle), and intervention (right), models.

Table 1 | Predicting trials to criterion at 12 months.

12 months trials to criterion								
Step	Predictors	R ²	R ² change	F change	p	β	Beta	p
1	Model group	0.39	0.39	17.88	< 0.01	1.67	0.62	< 0.01
2	5 months criterion	0.57	0.19	5.64	0.01	-0.24	-0.26	0.46
	7 months criterion					0.73	0.68	0.02

Our results show that the model can autonomously transform itself from a familiarity to novelty seeking model over development. As the model explored its virtual world, it accumulated traces in the HL. Over time, this experience helped the model quickly encode items and form stable WM peaks. This, in turn, enabled the model to actively represent known items and explore novel ones. Our results also showed that the initial conditions of the model created differences in the familiarity-to-novelty shift like those observed between term and preterm infants (Rose et al., 2002; see also Rose et al., 2001). Specifically, when we set the initial conditions of the preterm model to have weak neural interactions, the model shifted toward novelty more slowly over development, much like preterm infants do. Interestingly, we found that the experience the preterm infant model accumulated in the HL was comparable to the term infant model. This indicated that experience can create developmental change in the familiarity-to-novelty shift but the initial conditions play a major role in population differences.

Critically, these constraints are soft constraints: when we performed an intervention where we biased the model's pattern of looking, the developmental trajectory shifted in individual simulations. In particular, the intervention helped the models dwell on objects longer, creating stronger memory traces in the HL. In some models, this had advantageous effects: these models encoded items more quickly into WM and exhibited

novelty-seeking behaviors late in the first year that mimicked the pattern of term infants. In other models, however, the Hebbian traces in the perceptual field became too strong and the models showed a developmental regression with a bias toward familiarity.

The large variability in the outcomes of the intervention models is consistent with recent intervention studies that have trained caregivers to maintain their infants' attentional focus on objects. These interventions have facilitated positive developmental outcomes for children in areas of language, coordinated joint attention, and increased frequency with which caregivers maintain attentional focus (Landry et al., 2008). Nevertheless, the impact of such intervention studies has been diluted by individual differences in infants and caregivers. For example, preterm infants who experienced severe neonatal complications do not benefit from caregiver responsiveness to the same degree as infants who experienced relatively less severe neonatal complications (Landry et al., 2006). To optimize intervention, then, we need to tailor intervention to individuals.

We suggest that the DNF model might be a useful tool in these efforts. For instance, our simulation results suggest that we could initialize models to capture the performance of very young preterm infants in standard laboratory tasks. Critically, we could initialize models to capture the performance of individuals, not simply groups. We could then simulate different long-term interventions with these models and observe the predicted outcomes.

This could help researchers design individualized intervention regimens. Importantly, the models not only predict long-term outcomes, but also short-term benchmarks in performance. For instance, we could assess the models and infants at 3 months intervals in standard laboratory tasks to determine whether infants' looking and learning abilities match what is predicted by each infant's model. This provides multiple benchmarks to determine whether the intervention is on track.

Although the work presented here suggests that the DNF model could be a useful intervention tool, achieving this vision will require multiple layers of innovation. Most critically, the intervention we implemented was overly simplistic and ignores a fundamental factor in development—the role of other agents in infants' cognitive development. Infants develop in a rich social context that involves other agents such as parents and siblings. As described above, how other agents interact with infants while exploring objects can have a profound impact on how infants distribute their looks in time and in space as well as social interactions between infants and their caregivers (Perrinello and Ruff, 1988; see also Landry and Chaireskie, 1988). We are currently probing how a dyadic system that consists of parent and infant models sharing the same environment might explain the role of individual differences in parents and infants on the outcome of interventions as well as the emergence of social interactions in exploratory settings. This advancement will open the door to probe optimal intervention conditions for each parent-infant dyad. This may have far reaching practical implications.

Using the DNF model as an intervention tool in future work will also require tackling several challenges we simplified in the present simulation experiments. Conceptually, our model developed over the course of months. In practice, however, we simulated the model for much less time. This reflected the goals of this paper—to examine whether it was possible to have an autonomous model develop its own transition in visual exploratory biases. But using the model in more practical applications such as designing interventions will require that we more closely approximate the real-world experience of individual infants. We also encountered several practical challenges in the simulations that will be even more dramatic in more realistic simulation efforts. For instance, sometimes our models showed overly robust WM peaks that would endure for long periods of time. This would create a strong Hebbian trace that could dominate the looking and learning dynamics. We prevented this, in part, by carving the simulations up into episodes and re-initializing the layers every 10,000 time steps. In a more realistic model, we suspect this could be handled by adding more noise sources. For instance, data with infants suggests that their attentional abilities wax and wane over time (Oakes and Ross-Sheehy, 2004). We could implement this type of attentional inertia by adding a noisy resting level to the WM and PF layers that would gradually raise and lower slowly over time. The troughs in this type of attention would de-stabilize even robust WM peaks. This suggests that noise could serve an adaptive function in early development, facilitating exploration and ensuring that the system does not get stuck focusing too much on one thing.

This brings us back to the central issue we started with: what motivates infants to move from an initial bias toward familiarity

to a robust bias toward novelty? In one sense, our simulations suggest that there is no motivational source that propels the system forward in development. The DNF model propels itself forward because it is a complex, exploratory, dynamical system that accumulates its own history over time. Each time the DNF model formed a WM peak, this neural event left a trace in HLWM. The accumulation of this history over time raised the overall excitability of the WM field, leading to more robust WM peaks and the active maintenance of familiar items. This qualitatively new cognitive ability enabled the model to actively recognize what is known and explore new items in the environment. Thus, our autonomously developing model shows how changes in infants' visual exploratory skill measured in laboratory tasks can emerge from the accumulation of experience outside of the lab. There is no special motivating force that propels the model forward through development; rather, exploration and skill development come "for free" given the complex, self-organizing neural dynamics of the visual exploratory system. This is nicely illustrated by the full range of simulations we reported. Not all of our simulations developed a novelty bias—at least one of the intervention simulations showed a developmental regression, returning to familiarity-seeking behavior.

We contend that exploration is a fundamental, emergent property of complex dynamical systems—such systems can't help but explore (Thelen and Smith, 1994). In particular, given the high-dimensional nature of coupled behavioral and neural systems, such systems are inherently variable as they exchange energy with the surrounds and pass activation back-and-forth among different components of the system (Kelso, 1995). Such systems are also self-organizing, routinely settling in temporarily stable organizational states. Exploration emerges from the inherent tension between stability and variability. And in high-dimensional systems, this tension inevitably leads to new possible patterns of organization. Critically, complex dynamical systems are also historical, carrying this history forward through time. This sets the stage for new organizational patterns to be continually revisited and re-evaluated. Selection of adaptive states can then occur (Edelman, 1987).

There is another sense, however, in which our simulations suggest a motivational source is at work as infants transition from familiarity- to novelty-seeking. Oudeyer and Kaplan (2007) proposed two characterizations of intrinsic motivation. The first was a force that propels development forward, the notion of intrinsic motivation that is common in developmental psychology. As described above, this source was seemingly absent from the DNF model as it transitioned from familiarity- to novelty-seeking. The second characterization was in terms of the neurocognitive mechanisms that drive action. Conceptually, the idea is that subjective experiences of interestingness, ambiguity, and surprise move one to act. These subjective experiences might be driven by several neurocognitive mechanisms. Interestingness, for instance, can be driven by the degree to which an expected and experienced outcome differs. This sense of intrinsic motivation is present in the DNF model. Specifically, the pattern of connectivity among the layers of excitatory and inhibitory neurons in the model implements a neurocognitive mechanism that can identify "known"

from “unknown”—“expected” from “unexpected”—and then drive exploratory behavior.

If intrinsic motivation is inherent in the architecture of the model, a central question is where this architecture comes from. In our simulations, the architecture is assumed to be present early in development. Data are consistent with this conjecture. For instance, newborns exhibit evidence of recognizing stimuli experienced prenatally (DeCasper and Spence, 1986). But such data merely shifts the question of origins earlier. In our view, the neural architecture we proposed is likely a result of early prenatal developmental processes that are heavily dependent on patterned neural activity. For instance, recent work suggests that the type of neural architecture used here—DNFs—can emerge from a self-organizing process (Alecú et al., 2011; Detorakis and Rougier, 2012). Thus, the type of connectivity we assumed does not have to be “hard wired” in any sense—it can emerge during the course of early brain development. This also suggests that the neural system we proposed might be ubiquitous across species, consistent with evidence showing novelty-seeking behaviors in rabbits (Smith and Litvaitis, 2000), birds (Blough, 1984), and squirrels (Duncan and Jenkins, 1998).

In this context, it is important to note that novelty-seeking might not be the only outcome of autonomous visual exploration. In some studies, infants, and even adults, seek familiarity for items they do in fact have a robust memory for (Dodd et al., 2009). Seeking familiarity is clearly valuable in achieving practical goals—we often search for our coffee mug, keys, and so on. We are currently exploring how the DNF model might organize itself as a familiarity-seeking model in some contexts and novelty-seeking model in others.

REFERENCES

- Adolph, K. E., Cole, W. G., Komati, M., Garciaurre, J. S., Badaly, D., Lingeman, J. M., et al. (2012). How do you learn to walk: thousands of steps and dozens of falls per day. *Psychol. Sci.* 23, 1387–1394. doi: 10.1177/0956797612446346
- Adolph, K. E., and Robinson, S. R. (2013). “The road to walking: what learning to walk tells us about development,” in *Oxford Handbook of Developmental Psychology*, Vol. 1, ed P. Zelazo (New York, NY: Oxford University Press), 403–443.
- Alecú, L., Frezza-Buet, H., and Alexandre, F. (2011). Can self-organization emerge through dynamic neural fields computation. *Conn. Sci.* 23, 1–31. doi: 10.1080/09540091.2010.526194
- Blough, P. M. (1984). Visual search in pigeons: effects of memory set size and display variables. *Percept. Psychophys.* 35, 344–352. doi: 10.3758/BF03206338
- Cohen, L. B. (1972a). Attention-getting and attention-holding processes of infant visual preferences. *Child Dev.* 43, 869–879. doi: 10.2307/1127638
- Cohen, L. B. (1972b). “A two process model of infant visual attention,” in *Paper presented at the Merrill Palmer Conference on Research and Teaching of Infancy Development* (Detroit, MI).
- Colombo, J. (1995). On the neural mechanisms underlying developmental and individual differences in infant fixation duration: two hypotheses. *Dev. Rev.* 15, 97–135. doi: 10.1006/drev.1995.1005
- Colombo, J., and Mitchell, D. W. (1990). “Individual differences in early visual attention: fixation time and information processing,” in *Individual Differences in Infancy: Reliability, Stability, and Prediction*, eds J. Colombo and J. Fagen (Hillsdale, NJ: Lawrence Erlbaum Associates), 193–227.
- Colombo, J., Mitchell, D. W., O’Brien, M., and Horowitz, F. D. (1987). Stability of infant visual habituation across the first year. *Child Dev.* 58, 474–487. doi: 10.2307/1130524
- DeCasper, A. J., and Spence, M. J. (1986). Prenatal maternal speech influences newborns’ perception of speech sounds. *Infant Behav. Dev.* 9, 133–150. doi: 10.1016/0163-6383(86)90025-1
- Detorakis, G. I., and Rougier, N. P. (2012). A neural field model of the somatosensory cortex: formation, maintenance and reorganization of ordered topographic maps. *PLoS ONE* 7:e40257. doi: 10.1371/journal.pone.0040257
- Dodd, M. D., Van der Stigchel, S., and Hollingworth, A. (2009). Novelty is not always the best policy: inhibition of return and facilitation of return as a function of visual task. *Psychol. Sci.* 20, 333–339. doi: 10.1111/j.1467-9280.2009.02294.x
- Duncan, R. D., and Jenkins, S. H. (1998). Use of visual cues in foraging by a diurnal herbivore, Belding’s ground squirrel. *Can. J. Zool.* 76, 1766–1770. doi: 10.1139/z98-119
- Edelman, G. M. (1987). *Neuronal Darwinism: The theory of neuronal group selection*. New York, NY: Basic Books.
- Fantz, R. L. (1974). Visual experience in infants: decreased attention to familiar patterns relative to novel ones. *Science* 146, 668–670. doi: 10.1126/science.146.3644.668
- Fisher-Thompson, D., and Peterson, J. A. (2004). Infant side biases and familiarity-novelty preferences during a serial paired-comparison task. *Infancy* 5, 309–340. doi: 10.1207/s15327078in0503_4
- Gibson, E. J. (1988). Exploratory behavior in the development of perceiving, acting, and the acquiring of knowledge. *Annu. Rev. Psychol.* 39, 1–31. doi: 10.1146/annurev.ps.39.020188.000245
- Hunter, M. A., and Ames, E. W. (1988). “A multifactor model of infant preferences for novel and familiar stimuli,” in *Advances in Infancy Research*, Vol. 5, eds C. Rovee-Collier and L. O. Lipsitt (Norwood, NJ: Ablex), 69–95.
- Kelso, J. A. (1995). *Dynamic Patterns: The Self-Organization of Brain and Behavior*. Cambridge, MA: MIT Press.
- Kovack-Lesh, K. A., Horst, J. S., and Oakes, L. M. (2008). The cat is out of the bag: previous experience and online comparison jointly influence infant categorization. *Infancy* 13, 285–307. doi: 10.1080/15250000802189428

In summary, a robust developmental trend in infants’ visual exploration is that infants transition away from familiarity and toward novelty. This trend has largely been described as a by-product of faster processing speed; as processing speed increases, new items become familiar more quickly to infants and they are free to explore novelty. Our simulations indicate that novelty seeking and processing speed mutually support the development of each other. As infants explore more items along a dimension, they become increasingly familiar with that dimension. This, in turn, enables them to quickly form memories for items on that dimension and continue to explore novelty. We gained this insight by using a DNF model of infant visual exploration to ask what motivates an infant to switch their exploratory style from familiarity- to novelty-seeking. The DNF model propelled itself forward simply by autonomously accumulating a learning history as it explored a virtual visual world with a reasonable degree of stimulus variation. In this sense, no motivational force was required for the model to shift its exploratory style. In another sense, however, the pattern of neuronal connectivity in the model clearly sets the stage for this shift to happen. Most critically, our simulations suggest that the accumulation of real-time exploratory behavior is powerful enough to create developmental change.

ACKNOWLEDGMENTS

Preparation of this manuscript was supported by R01MH62480 awarded to John P. Spencer. The writing this manuscript was also supported by R01HD045713 awarded to Larissa K. Samuelson. The funding sources had a role in the design and conclusions drawn from the research presented here.

- Landry, S. H., and Chapeskie, M. L. (1988). Visual attention during toy exploration in preterm infants: effects of medical risk and maternal interactions. *Infant Behav. Dev.* 11, 187–204. doi: 10.1016/S0163-6383(88)80005-5
- Landry, S. H., Smith, K. E., and Swank, P. R. (2006). Responsive parenting: establishing early foundations for social, communication, and independent problem-solving skills. *Dev. Psychol.* 42, 627–642. doi: 10.1037/0012-1649.42.4.627
- Landry, S. H., Smith, K. E., Swank, P. R., and Guttentag, C. (2008). A responsive parenting intervention: the optimal timing across early childhood for impacting maternal behaviors and child outcomes. *Dev. Psychol.* 44, 1335–1353. doi: 10.1037/a0013030
- Oakes, L. M., Horst, J. S., Kovack-Lesh, K. A., and Perone, S. (2008). “How infants learn categories,” in *Learning and The Infant Mind*, eds A. Woodward and A. Needham (New York, NY: Oxford University Press), 144–171. doi: 10.1093/acprof:oso/9780195301151.003.0006
- Oakes, L. M., and Ross-Sheehy, S. (2004). Attentional engagement in infancy: the interactive influence of attentional inertia and attentional state. *Infancy* 5, 239–252. doi: 10.1207/s15327078in0502_8
- Oudeyer, P.-Y., and Kaplan, F. (2007). What is intrinsic motivation. A typology of computational approaches. *Front. Neurorobot.* 1:6. doi: 10.3389/neuro.12.006.2007
- Perone, S., Simmering, V. R., and Spencer, J. P. (2011). Stronger neural dynamics capture developmental changes in infants’ visual working memory capacity over development. *Dev. Sci.* 14, 1379–1392. doi: 10.1111/j.1467-7687.2011.01083.x
- Perone, S., and Spencer, J. P. (2013a). Autonomy in action: linking the act of looking to memory formation in infancy via dynamic neural fields. *Cogn. Sci.* 37, 1–60. doi: 10.1111/cogs.12010
- Perone, S., and Spencer, J. P. (2013b). The co-development of looking dynamics and discrimination performance. *Dev. Psychol.* doi: 10.1037/a0034137. [Epub ahead of print].
- Perrinello, R. M., and Ruff, H. A. (1988). The influence of adult intervention on infants’ level of attention. *Child Dev.* 59, 1125–1135. doi: 10.2307/1130279
- Piaget, J. (1952). *The Origins of Intelligence in Children*. New York, NY: International Universities Press, Inc. doi: 10.1037/11494-000
- Quinn, P. C., Yahr, J., Kuhn, A., Slater, A. M., and Pascalis, O. (2002). Representation of the gender of human infants: a preference for females. *Perception* 31, 1109–1121. doi: 10.1080/p3331
- Robinson, C. W., and Sloutsky, V. M. (2007). Visual processing speed: effects of auditory input on visual processing. *Dev. Sci.* 10, 734–740. doi: 10.1111/j.1467-7687.2007.00627.x
- Roder, B. J., Bushnell, E. W., and Saserville, A. M. (2000). Infants’ preferences for familiarity and novelty during the course of visual processing. *Infancy* 1, 491–507. doi: 10.1207/S15327078IN0104_9
- Rose, S. A., Gottfried, A. W., Melloy-Carminar, P. M., and Bridger, W. H. (1982). Familiarity and novelty preferences in infant recognition memory: implications for information processing. *Dev. Psychol.* 18, 704–713. doi: 10.1037/0012-1649.18.5.704
- Rose, S. A., Feldman, J. F., and Jankowski, J. J. (2001). Attention and recognition memory in the 1st year of life: a longitudinal study of preterm and full-term infants. *Dev. Psychol.* 37, 135–151. doi: 10.1037/0012-1649.37.1.135
- Rose, S. A., Feldman, J. F., and Jankowski, J. J. (2002). Processing speed in the 1st year of life: a longitudinal study of preterm and full-term infants. *Dev. Psychol.* 38, 895–902. doi: 10.1037/0012-1649.38.6.895
- Rose, S. A., Feldman, J. F., and Jankowski, J. J. (2004). Infant visual recognition memory. *Dev. Rev.* 24, 74–100. doi: 10.1016/j.dr.2003.09.004
- Rose, S. A., Feldman, J. F., and Jankowski, J. J. (2007). “Developmental aspects of visual recognition memory in infancy,” in *Short- and Long-Term Memory in Infancy and Early Childhood*, eds L. M. Oakes and P. J. Bauer (New York, NY: Oxford University Press), 153–178.
- Rose, S. A., Feldman, J. F., and Jankowski, J. J. (2009). Information processing in toddlers: continuity from infancy and persistence of preterm deficits. *Intelligence* 37, 311–320. doi: 10.1016/j.intell.2009.02.002
- Rose, S. A., Feldman, J. F., and Jankowski, J. J. (2012). Implications of infant cognition for executive functions at age 11. *Psychol. Sci.* 23, 1345–1355. doi: 10.1177/0956797612444902
- Schutte, A. R., and Spencer, J. P. (2009). Tests of the dynamic field theory and the spatial precision hypothesis: capturing a qualitative developmental transition in spatial working memory. *J. Exp. Psychol. Hum. Percept. Perform.* 35, 1698–1725. doi: 10.1037/a0015794
- Schutte, A. R., Spencer, J. P., and Schoner, G. (2003). Testing the dynamic field theory: working memory for locations becomes more spatially precise over development. *Child Dev.* 74, 1393–1417. doi: 10.1111/1467-8624.00614
- Shinskey, J. L., and Munakata, Y. (2005). Familiarity breeds searching: Infants reverse their novelty preferences when reaching for hidden objects. *Psychol. Sci.* 16, 596–600. doi: 10.1111/j.1467-9280.2005.01581.x
- Simmering, V. R., Spencer, J. P., and Schutte, A. R. (2007). Generalizing the dynamic field theory of spatial cognition across real and developmental time scales. *Brain Res.* 1202C, 68–86. doi: 10.1016/j.brainres.2007.06.081
- Smith, D. F., and Litvaitis, J. A. (2000). Foraging strategies of sympatric lagomorphs: implications for differential success in fragmented landscapes. *Can. J. Zool.* 78, 2134–2141. doi: 10.1139/z00-160
- Thelen, E., and Smith, L. B. (1994). *A Dynamic Systems Approach to the Development of Cognition and Action*. Cambridge, MA: MIT Press.
- Wetherford, M. J., and Cohen, L. B. (1973). Developmental changes in infant visual preferences for novelty and familiarity. *Child Dev.* 44, 416–424. doi: 10.2307/1127994

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 21 June 2013; accepted: 30 August 2013; published online: 20 September 2013.

Citation: Perone S and Spencer JP (2013) Autonomous visual exploration creates developmental change in familiarity and novelty seeking behaviors. Front. Psychol. 4:648. doi: 10.3389/fpsyg.2013.00648

This article was submitted to Cognitive Science, a section of the journal Frontiers in Psychology.

Copyright © 2013 Perone and Spencer. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

APPENDIX

The notation used in the equations is presented in **Table A1**.

MODEL EQUATIONS

Each neuronal layer is specified by a differential equation numerically integrated using the Euler method.

Perceptual field (PF)

PF consists of reciprocally coupled excitatory, $PF(u)$, and inhibitory, $Inhib(v)$, layers for dimension x . The excitatory layer of PF is given by the following equation:

$$\begin{aligned} \tau_e \dot{u}(x, t) = & -u(x, t) + h_u \\ & + a_{ul} \sum_{l=1}^n g(l_i) + \sum_{i=1}^n s_i(x, t)g(l_i) \\ & + \int c_{uu}(x - x')g(u(x', t))dx' \\ & - \int c_{uv}(x - x')g(v(x', t))dx' \\ & - a_{uv_global} \int g(v(x', t))dx' \\ & + \int c_{um}(x - x')m(x', t)dx' \\ & + \int c_r(x - x')\xi(x', t)dx' \end{aligned}$$

where $\dot{u}(x, t)$ is the rate of change of activation in the excitatory layer of PF across the continuous behavioral dimension, x , as a function of time, t . τ_e is the time constant along which excitatory activation evolves. Activation within PF is influenced by its current state, $u(x, t)$, and its negative neuronal resting level, h_u . PF

Table A1 | Notation.

Letter	Meaning
a	Amplitude/strength parameter
x, y	Dimension (x = color, y = shape)
l_i	Looking nodes (i = index of the node)
u	Activation variable for PF
v	Activation variable for Inhib
w	Activation variable for WM
m	Activation variable for memory/Hebbian layer
s	Stimulus input (Gaussian for fields)
c	Connection weight function
g	Gating function
t	Time
τ	Time scale parameter
h	Resting level (static or dynamic)
n	Number of nodes
r	Random contribution
ξ	Noise parameter
e	Excitatory
i	Inhibitory

receives a global boost from the fixation system, $a_{ul} \sum_{l=1}^n g(l_i)$, which is dictated by the gating function, $g(l_i)$, and weighted by the amplitude or “strength” parameter, a_{ul} . This means that when a task-relevant location is fixated, PF receives a boost of activation. PF also receives stimulus input at the suprathreshold fixated location, $\sum_{l=1}^n s_i(x, t)g(l_i)$, where $s_i(x, t)$ is a Gaussian input (see below) distributed across the behavioral dimension, x . Note that for these inputs $n = 2$ because only looking nodes associated with the left and right locations are associated with task-relevant stimuli in the task space (see “Fixation System” below).

The gating function is given by the following equation which takes a sigmoidal shape over the activation variable, u :

$$g(u) = \left[\frac{1}{1 + \exp[-\beta(u(t) - u_0)]} \right],$$

where β is the slope of the sigmoid function and u_0 is the threshold (0).

The stimulus input takes the form of a Gaussian distributed over the behavioral dimension, x :

$$s(x, t) = a \exp \left[-\frac{(x - \mu)^2}{2\sigma^2} \right] \chi(t)$$

with stimulus position centered at μ , strength a (set to 17), and width σ (set to 3). The gating function, $\chi(t)$, is set to 1 when the stimulus is present and 0 otherwise.

PF dynamics are also influenced by local excitatory within-layer interactions, $\int c_{uu}(x - x')g(u(x', t))dx'$. These interactions are specified by the convolution of a Gaussian profile, $c_{uu}(x - x')$, which determines the neighborhood across which excitatory interactions propagate and a non-linear gating function, $g(u(x', t))dx'$, dictating that only neurons with above threshold activation (>0) participate in the interactions.

The Gaussian convolution was defined by:

$$c(x - x') = a \exp \left[-\frac{(x - x')^2}{2\sigma^2} \right]$$

where a sets the amplitude and σ sets the width (i.e., standard deviation) of the connection matrix function.

PF dynamics are also influenced by two inhibitory components. The first is a local inhibitory component, $\int c_{uv}(x - x')g(v(x', t))dx'$. Inhibitory interactions are projected across a neural neighborhood specified by a Gaussian, $c_{uv}(x - x')$, and only above-threshold activity in the inhibitory layer contribute to interactions. The second is a global inhibitory component, $a_{uv_global} \int g(v(x', t))dx'$, where the sum of suprathreshold activity within the inhibitory layer across the behavioral dimension, x , at time, t , is weighted by a_{uv_global} .

The last contribution to PF dynamics is spatially correlated noise, which is presented to PF by convolving a field of white noise with a Gaussian kernel, $\int c_r(x - x')\xi(x', t)dx'$, with strength, a_r , set to 0.025 and width, σ_r , set to 1.

Inhibitory field (Inhib)

The excitatory layers PF(u) and WM(w) are reciprocally coupled to an inhibitory layer, Inhib(v). The equation for Inhib is:

$$\begin{aligned}\tau_i \dot{v}(x, t) = & -v(x, t) + h_v \\ & + \int c_{vu}(x - x')g(u(x', t))dx' \\ & + \int c_{vw}(x - x')g(w(x', t))dx' \\ & + \int c_r(x - x')\xi(x', t)dx'\end{aligned}$$

where $\dot{v}(x, t)$ specifies the rate of change of activation for each neuron along the behavioral dimension, x , as a function of time, t . τ_i is the time constant along which inhibitory activation evolves. Activation in Inhib is influenced by its current state, $v(x, t)$, and its resting level, h_v . Inhib receives excitatory inputs from PF, $\int c_{vu}(x - x')g(u(x', t))dx'$, and WM, $\int c_{vw}(x - x')g(w(x', t))dx'$. These inputs are projected across a neural neighborhood specified by a Gaussian projection, $c(x - x')$, to which only suprathreshold neurons in PF and WM contribute as dictated by the gating function, g . An independent source of spatially correlated noise is also added to the inhibitory layer, $\int c_r(x - x')\xi(x', t)dx'$.

Working memory field (WM)

The WM(w) field is given by the following equation:

$$\begin{aligned}\tau_e \dot{w}(x, t) = & -w(x, t) + h_w \\ & + a_{ws} \sum_{l=1}^n s_l(x, t)g(l_i) \\ & + \int c_{wu}(x - x')g(u(x', t))dx' \\ & + \int c_{ww}(x - x')g(w(x', t))dx' \\ & + \int c_{vv}(x - x')g(v(x', t))dx' \\ & - a_{wv_global} \int g(v(x', t))dx' \\ & + \int c_{wm}(x - x')m(x', t)dx' \\ & + \int c_r(x - x')\xi(x', t)dx'\end{aligned}$$

The equation for WM is identical to the equation for PF with two exceptions. First, the input from the fixation system differs: there is no global boost in activation from the fixation system into WM, and the stimulus input to WM, $a_{ws} \sum_{l=1}^n s_l(x, t)g(l_i)$, is weighted by a strength parameter, a_{ws} , which was set to .05. Second, WM receives an excitatory input from PF, $\int c_{wu}(x - x')g(u(x', t))dx'$.

Memory/Hebbian layers (HL)

Activation in PF and WM is influenced by traces in an associated memory (m) or Hebbian layer (HL), which implement a form

of Hebbian learning (see text). The equations for each HL are identical. The equation for the HL associated with PF is:

$$\dot{m}_u(x, t) = \begin{cases} (-m_u(x, t) + g(u(x, t))/\tau_{m_build}) & \text{if } u(x, t) > 0 \\ (-m_u(x, t))/\tau_{m_decay} & \text{otherwise} \end{cases}$$

where $\dot{m}_u(x, t)$ is the rate of change of activation for each site, x , in HL as a function of time, t . The constants τ_{m_build} and τ_{m_decay} set the time scale along which activation traces accrue and decay, respectively. Activation in HL only accrues when there is suprathreshold activation in PF. Otherwise, activation in HL decays.

Fixation system

The fixation system consists of four nodes that stochastically look at left and right locations (at which stimuli can appear) and center and away locations (at which no task-relevant stimuli appear). The nodes interact in a mutually inhibitory, winner-takes-all fashion. The equation for the fixation system is:

$$\begin{aligned}\tau_e \dot{l}_i(t) = & -l_i + h_i(t) + s_i(t) \\ & + a_{iig}(l_i) \\ & + a_{lu}g(l_i) \int g(u(x', t))dx' \\ & - a_{l_global} \sum_{j \neq i} g(l_j)\end{aligned}$$

where the activation variable, l , is set by the excitatory time scale, τ_e . Activation of each looking node is influenced by its current state, l , and its dynamic negative resting level, $h_i(t)$ (described below). Activation of each looking node is also influenced by a stimulus input given by:

$$s_i(t) = a_{i_tonic}(t)(a_i + \xi(t)) + a_{i_transient}(t)$$

The stimulus associated with each node is different (see “Fixation System Parameters” below) to reflect the different stimulus properties of the attention-getter at the central location, the stimuli at the left and right locations, and non-task-relevant input at all “away” locations. The left and right nodes are presented with a noisy input at each time step when a stimulus is present, $a_{i_tonic}(t)(a_i + \xi(t))$, and a transient input to signify the appearance of a stimulus, $a_{i_transient}(t)$, present for the initial 75 time steps of each stimulus presentation. The away node is continuously presented with a noisy input to signify the “tonic” presence of stimuli in the task space. The center node is presented only with a transient input to reflect attention-getting stimuli briefly present at the onset of a trial (in our simulations, 50 time steps), effectively driving the fixation system to switch gaze from the away location to the center location.

The gating function, g , dictates the presence of a self-excitatory component to each looking node, $a_{iig}(l_i)$, and the passing of a negative, inhibitory input to all other nodes, $a_{l_global} \sum_{j \neq i} g(l_j)$, with weight a_{l_global} . The gating function also regulates the presence of input to the fixation system from the perceptual field

Table A2 | Neurocognitive system parameters.

PF(<i>u</i>)			WM(<i>w</i>)			Inhib(<i>v</i>)			Time scales (τ)		Memory layers (M)	
Term	Preterm	Term	Preterm	Term	Preterm	Term	Preterm					
h_u	-10	-	h_w	-5	-	h_v	-10	-	τ_e	80	c_{um}	2.5
a_{uu}	0.75	0.6	a_{ww}	2.0075	1.606	a_{uv}	0.459	0.3672	τ_i	10	σ_{um}	3
σ_{uu}			σ_{ww}	3	-	σ_{uv}	10	-	τ_{build}	20,000	c_{wm}	1.5
			σ_{wu}	1.2	-	σ_{vu}	0.2	-	τ_{decay}	400,000	σ_{wm}	5
			σ_{wu}	5	-	σ_{vu}	5	-	τ_h	80		
			σ_{vw}	4.5	-	a_{ww}	0.405	0.324				
			σ_{vw}	5	-	σ_{vv}	30	-				
						a_{vv}	4.5	3.6				
						σ_{vv}	5	-				
						a_{wv_global}	0.01	-				
						a_{uv_global}	0	-				

Table A3 | Fixation system parameters.

	Location			
	Left	Right	Center	Away
a_{l_global}	1.8	-	-	-
a_{ii}	2.00	-	-	-
a_{lu}	0.25	-	-	-
a_{ui}	1.00	-	-	-
$a_{i_transient}$	3.00	3.00	15	0
a_{i_tonic}	5.60	-	-	-
a_i	0.70	-	-	-
a_{h_rest}	-5.00	-	-	-
a_{h_down}	-3.60	-	-	-

across dimension x , $a_{lug}(l_i) \int g(u(x', t)) dx'$, with weight a_{lu} . Note that these inputs are set to 0 for the looking nodes associated with the center and away locations because there is no stimulus presented at those locations.

The resting level of each looking node is dynamic and is governed by the following equation:

$$\tau_h \dot{h}_i(t) = -h_i(t) + a_{h_rest} + a_{h_low}g(l_i)$$

where τ_h sets the time scale along which the resting level of each node, h_i , evolves. When the current level of activation of a looking node is above threshold [determined by the gating function, $g(l_i)$] the resting level decreases toward a low attractor, the sum of a_{h_rest} and a_{h_low} (which are both negative values). When the current level of activation of a looking node is below threshold, the resting level returns to baseline, a_{h_rest} .

MODEL PARAMETERS

Table A2 shows the parameters for the neurocognitive system and **Table A3** shows the parameters for the fixation system used to simulate the looking behavior of term and preterm infants. To create the preterm infant model, we began with the term infant parameters and manipulated the parameters used to implement the SPH (see Schutte and Spencer, 2009; see also Perone et al., 2011; Perone and Spencer, 2013a,b). This involved uniformly decreasing the strength of within-layer excitatory connections in PF (a_{uu}) and WM (a_{ww}) and across layer inhibitory connections from inhib to PF (a_{uv}) and to WM (a_{vv}) by 20%. The SPH parameters are shown in bold. All other parameters were fixed for the term and preterm models. Note that for the intervention simulations (see text), 0.0625 was added to a_i when left or right node was suprathreshold.



Image free-viewing as intrinsically-motivated exploration: estimating the learnability of center-of-gaze image samples in infants and adults

Matthew Schlesinger^{1*} and Dima Amso²

¹ Department of Psychology, Southern Illinois University, Carbondale, IL, USA

² Cognitive, Linguistic, and Psychological Sciences, Brown University, Providence, RI, USA

Edited by:

Gianluca Baldassarre, Italian National Research Council, Italy

Reviewed by:

Sufen Chen, Albert Einstein College of Medicine, USA

Daniele Caligiore, Institute of Cognitive Sciences and Technologies, Italy

Martin Thirkettle, The Open University, UK

Valerio Sperati, Consiglio Nazionale delle Ricerche, Italy

***Correspondence:**

Matthew Schlesinger, Department of Psychology, Southern Illinois University, Life Science II, Rm. 281, Carbondale, IL 62901, USA
e-mail: matthews@siu.edu

We propose that free viewing of natural images in human infants can be understood and analyzed as the product of intrinsically-motivated visual exploration. We examined this idea by first generating five sets of center-of-gaze (COG) image samples, which were derived by presenting a series of natural images to groups of both real observers (i.e., 9-month-olds and adults) and artificial observers (i.e., an image-saliency model, an image-entropy model, and a random-gaze model). In order to assess the sequential learnability of the COG samples, we paired each group of samples with a simple recurrent network, which was trained to reproduce the corresponding sequence of COG samples. We then asked whether an intrinsically-motivated artificial agent would learn to identify the most successful network. In Simulation 1, the agent was rewarded for selecting the observer group and network with the lowest prediction errors, while in Simulation 2 the agent was rewarded for selecting the observer group and network with the largest rate of improvement. Our prediction was that if visual exploration in infants is intrinsically-motivated—and more specifically, the goal of exploration is to learn to produce sequentially-predictable gaze patterns—then the agent would show a preference for the COG samples produced by the infants over the other four observer groups. The results from both simulations supported our prediction. We conclude by highlighting the implications of our approach for understanding visual development in infants, and discussing how the model can be elaborated and improved.

Keywords: visual exploration, perceptual development, intrinsic motivation, eye movements, image free-viewing

INTRODUCTION

Within minutes of birth, human infants open their eyes and begin to explore the visual world (Slater, 2002). Although neonates lack visuomotor experience—and their visual acuity is poor—their eye movements are not random (Fantz, 1956; Haith, 1980). Instead, infants' gaze patterns are organized in a manner that facilitates the discovery and learning of relevant visual features and objects, such as the caretaker's face (e.g., Maurer and Barrera, 1981; Bushnell et al., 1989; Morton and Johnson, 1991).

With additional experience, infants not only gain further control over their eye movements, but their gaze patterns also continue to develop. For example, during the first month after birth, infants tend to limit their scanning to a small portion of an image (Bronson, 1982, 1991). By age 3 months, however, infants produce gaze patterns that are more systematically distributed over visual scenes. During the same age period, comparable changes also occur in a number of other related visual skills, such as maintaining fixation of a target object in the presence of distracting stimuli, as well as selecting informative regions of the visual scene to fixate and encode (e.g., Johnson et al., 2004; Amso and Johnson, 2005).

There have been several important advances in the study of infants' gaze patterns. One approach leverages the tendency for

infants to orient toward salient, predictable events, and in particular, events that are contingent on infants' own actions (e.g., Haith et al., 1988; Kenward, 2010). For example, Wang et al. (2012) recently developed a gaze-contingent paradigm in which infants quickly learned to anticipate the appearance of a picture that was "triggered" by first fixating an object at another location. This work highlights the fact that infants' visual-activity is prospective and future-oriented.

A second advance is the use of image free-viewing methods, which record and analyze infants' eye movements as they view a series of images or video clips, often including naturalistic scenes (e.g., Aslin, 2009; Frank et al., 2009, 2012). In contrast to methods that present an implicit task to the infant, such as comparing two images or locating a target object, image free-viewing is comparatively less-constrained, and may more accurately reflect not only infants' spontaneous gaze patterns, but also the process of information pickup and learning that occurs in real time during visual exploration. While early work using image-free viewing tended to rely on somewhat coarse analytical methods, such as comparing time spent viewing specific regions of interest (ROIs; e.g., Bronson, 1982, 1991), more recent work in this area has employed relatively sophisticated quantitative methods. For example, Frank et al. (2009) computed the frame-by-frame image saliency of a

short animation clip (i.e., “A Charlie Brown Christmas”), and then compared infants’ attention to faces in the clip vs. their attention to high-salience non-face regions. A key finding from their analysis was that at age 3-months, infants’ gaze patterns were more strongly influenced by salience than by social stimuli such as faces; however, by age 9 months, this pattern reversed, and infants oriented reliably to faces.

Finally, the approach we propose here represents a third advance. In particular, there are several recent models that successfully capture the kinematic properties of infants’ gaze patterns during conventional tasks, such as preferential looking, gaze following, and visual search (e.g., Schlesinger et al., 2007; Triesch et al., 2007; Perone and Spencer, 2013). However, to our knowledge, our model is the first attempt to apply incremental, adaptive-learning methods (i.e., artificial neural networks and reinforcement learning) as a computational tool for analyzing infants’ gaze patterns during image free-viewing.

Specifically, we propose that in addition to analyzing the spatial distribution and timing of infants’ gaze patterns, the *sequential content of their fixations during image free-viewing* may also provide an important source of information. In particular, the sequence of fixations produced by an observer can be interpreted as a series of high-resolution visual samples, each centered at the corresponding gaze point (i.e., center-of-gaze or COG samples; Dragoi and Sur, 2006; Mohammed et al., 2012). As a form of exploration in the visual modality, these COG samples are similar to the tactile data generated by structured hand and finger movements during haptic object exploration (i.e., exploratory procedures or EPs; Klatzky and Lederman, 1990), insofar as different sampling patterns are the result of different exploration strategies.

In this paper, we propose that infants’ gaze patterns during image free-viewing are a form of visual exploration, and that the sequential structure embedded within these patterns can be analyzed with the theoretical framework of *intrinsic motivation*. More specifically, we suggest that:

Learning objective 1: over the short term (i.e., real time), the goal of visual exploration is to accurately predict the content of the next fixation (i.e., the subsequent COG sample), given the current fixation together with the history of recent fixations.

Learning objective 2: superimposed on the timescale of learning objective 1, a longer-term goal of visual exploration is to learn how to generate sequentially learnable gaze patterns, that is, to learn how to scan images or scenes such that the resulting set of COG samples is sequentially predictable.

Learning objective 1 is predicated on the idea that prediction-learning and future-oriented actions are pervasive characteristics of infant development (e.g., Haith, 1994; Johnson et al., 2003; von Hofsten, 2010). In addition, a related mechanism that may underlie prediction-learning is the detection of statistical patterns or regularities in the environment, such as those in linguistic input or natural scenes (e.g., Field, 1994; Saffran et al., 1996). However, a unique aspect of our proposal is that, rather than passively observing sensory patterns in the external world, infants may

also contribute to the process of pattern detection by embedding structure in their own exploratory behavior.

The rationale for learning objective 2, meanwhile, is that in addition to acquiring specific skills, such as learning to grasp or walk, infants also engage in behaviors that seem to have no explicit purpose, such as babbling or playing with blocks. In other words, *intrinsically-motivated* behaviors are done simply for the sake of learning (Oudeyer and Kaplan, 2007; Baldassarre and Mirolli, 2013; Schlesinger, 2013). This contrasts with *extrinsically-motivated* behaviors, which have a clear and (typically) biological benefit, such as obtaining food, rest, or sex (Baldassarre, 2011).

By this view, we argue that visual exploration serves two developmental functions. First, at the moment-to-moment level (learning objective 1), infants learn to discover and predict the particular statistical regularities of the images and scenes they are scanning (e.g., moving objects tend to remain on continuous trajectories, natural scenes are typically illuminated from above, “angry” eyes tend to co-occur with a frowning mouth, etc.). Second, and over a longer timescale (learning objective 2), infants are also “learning to learn,” that is, their scanning strategies are refined, and in particular, infants are improving in their ability to detect and attend to relevant visual features. In our model, we conceptualize this second-order learning process as an intrinsically-motivated artificial agent, which observes the performance of five scanning strategies, and is rewarded for selecting the strategy that produces the lowest (or most rapidly falling) prediction errors.

In order to pursue the first learning objective, we assigned five unique sets of COG samples to each of five simple recurrent networks (SRNs). We selected the SRN architecture as a computational tool for two specific reasons. First, it serves as a proxy for the statistical-learning mechanism noted above. In particular, it is well-suited to detecting regularities or statistical dependencies within temporal sequences of input. Second, we also exploited SRNs as a means to measure the relative predictability of the sequences produced by the observer groups. Specifically, the training errors produced by the SRN provide a straightforward metric for assessing learnability of the COG samples.

Each set of COG samples was generated by a different group of real or artificial observers: 9-month-olds, adults, an image-saliency model, an image-entropy model, and a random-gaze model. The task of each SRN is to learn to reproduce the sequence of COG samples produced by its corresponding group. We then pursued the second learning objective by creating an intrinsically-motivated artificial agent, which selects among the five SRNs as they are trained, and is rewarded for either selecting the SRN with the lowest errors (Simulation 1), or the SRN that learns the fastest (Simulation 2). We return to this issue below, where we describe the specific reward functions used to evaluate the choices of the intrinsically-motivated agent.

We reasoned that each group of real or artificial observers collectively represents a distinct scanning pattern or strategy, and as a result, the COG samples generated by each group should be differentially learnable. In addition, given our proposal that infants’ visual exploration is specifically geared toward the goals of (1) sequential predictability and (2) optimal prediction-learning, we

therefore, hypothesized that the COG samples produced by 9-month-olds would be selected first by an intrinsically-motivated agent, whether the reward function is based on learning errors (Simulation 1) or change in the rate of learning (Simulation 2). We also predicted that as reward diminishes in Simulation 2 (i.e., as learning of the infants' COG samples asymptotes), the agent should then shift its preference from the infants' COG samples to the adults' samples. This was an exploratory prediction, based on the assumption that adults' gaze patterns are not only influenced by sequential learnability (like infants), but that they are also informed by the observer's history of goal-directed activity (e.g., Shinoda et al., 2001; Hayhoe and Ballard, 2005).

The rest of the paper is organized as follows. We first describe the set of images presented to the five groups of observers, as well as the procedure used to acquire the gaze data from the human observers. We also describe the design of the three groups of artificial observers, and the analogous procedure used to generate the gaze data from each of these groups. We conclude this section by explaining how the gaze data were used to generate COG samples. In the next section, we then describe the architecture and learning algorithms used in the SRN prediction networks (PNs) and the intrinsically-motivated agent. Following this, we present Simulation 1, in which the artificial agent vicariously explores the COG samples by selecting among the five SRNs, and learns by trial-and-error to find the SRN with the lowest prediction errors. Next, in Simulation 2 we present the findings of a closely-related reward function, in which the agent is rewarded for finding the SRN with the fastest learning progress (i.e., the largest decline in the error rate over successive training epochs). In the final section, we relate our findings to the development of visual exploration in infants, and describe some ways to address the limitations of our current modeling approach.

MATERIALS

TEST IMAGES

Sixteen naturalistic, color images were used as stimuli for collecting eye movements, including 8 indoor and 8 outdoor scenes. One or more people were present in each image; in some images, the people were in the foreground, while in others they were in the background. **Figure 1** presents 4 of the 16 test images. The infant and adult observers were presented with the test images at the original image resolution (1680×1050 pixels), while the



FIGURE 1 | Four of the test images.

artificial observers were presented with downscaled versions of the images (480×300 pixels). As we note below, all of the infant and adult fixations were rescaled to the lower resolution, so that real and artificial observers' gaze data could be directly compared.

OBSERVER GROUPS

Real Observers

Eye-movement data were collected from 10 adults and 10 9-month-olds infants (mean ages = 19 years and 9.5 months, respectively). Except where noted, a comparable procedure was used for testing both adult and infant participants. All participants provided either signed consent for the study, or in the case of the infants, assent was provided by the infants' parents.

Participants sat about 70 cm from a 22" (55.9 cm) monitor. Infants sat in a parent's lap. Eye movements were recorded using a remote eye tracker (SMI SensoMotoric Instruments RED system). In addition, a standard digital video camera (Canon ZR960) was placed above the computer screen to record children's head movements. All calibration and task stimuli were presented using the Experiment Center software from SMI. Before beginning the task, point-of-gaze (POG) was calibrated by presenting an attractive, looming stimulus in the upper left and lower right corners of the screen. The same calibration stimulus was then presented in the four corners of the screen in order to validate the accuracy of the calibration.

We eye tracked participants as they freely scanned 16 color photographs depicting both indoor and outdoor scenes (see **Figure 1** for examples; for a comparable procedure, see also Amso et al., 2013). All images were presented for 5 s and spanned the entire display. The order of image presentation was randomized. A central fixation target was used to return participants' POG to the center of the screen between images.

Artificial Observers

The purpose of creating the artificial observers was to generate a set of synthetic gaze patterns, in which the underlying mechanism driving gaze from one location to the next was known in advance. In addition, the three groups of artificial observers also provide a well-defined baseline for comparison with the infant and adult observers (see Frank et al., 2009, for a similar approach).

Saliency model. The saliency model was designed to simulate an artificial observer whose gaze pattern is determined by bottom-up visual features, such as edges or regions with strong light/dark contrast. In particular, each test image was transformed by first creating three feature maps (tuned to oriented edges, luminance, and color contrast, respectively), and then summing the feature maps into a saliency map. We then used each saliency map to generate a series of simulated fixations.

- 1. Feature maps.** The original images were first downscaled to 480×300 . Next, each image was passed through a bank of image filters, resulting in three sets of feature maps: 4 oriented edge maps (i.e., tuned to 0° , 45° , 90° , and 135°), 1 luminance map, and 2 color-contrast maps (i.e., red-green and blue-yellow color opponency maps). In addition, this process

was performed over 3 spatial scales (i.e., to capture the presence of the corresponding features at high, medium, and low spatial frequencies), by successively blurring the original image and then repeating the filtering process [for detailed descriptions of the algorithms used for each filter type, refer to Itti et al. (1998) and Itti and Koch (2000)]. As a result, 21 total feature maps were computed for each test image.

2. **Saliency maps.** The saliency map was produced by first normalizing the 21 corresponding feature maps, and then summing them together. For the next step (simulating gaze data), each saliency map was downsampled to 48×30 . These resulting maps were then normalized, by dividing each map by the average of the highest 100 saliency values from that map. **Figure 2** illustrates the saliency map (left image) for one of the outdoor scenes (compare with the original image in **Figure 1**).
3. **Simulated gaze data.** In order to equate the mean number and frequency of gaze shifts across the real and artificial observers, the gaze data of the infants and adults were pooled, and the corresponding values were computed. This resulted in a mean of 13 fixations per image, and a mean latency of 300 ms between fixations. For the artificial observers, the simulated timestep was 33 ms per processing cycle (i.e., 30 updates per second). These values were then used as fixed parameters for the artificial observers. A single trial was simulated by iteratively updating a fixation map—which is the difference between the saliency map and a decaying inhibition map (see below)—and selecting a location on the fixation map every 300 ms. Note that the inhibition map served as an analog for an inhibition-of-return (IOR) mechanism, which allowed the saliency model to release its gaze from the current location and shift it to other locations on the fixation map.

Each trial began by selecting the initial fixation point at random. Next, the inhibition map was initialized to 0, and a 2D Gaussian surface was added to the map, centered at the current fixation point, with an activation peak equal to the value at the corresponding location on the saliency map. Over the subsequent 300 ms, activity on the inhibition map decayed at a rate of 10% per timestep. At 300 ms, the next fixation point was selected: (a) the fixation map was updated by subtracting the inhibition map from the saliency map (negative values were set to zero), (b) the top 100 values on the saliency map were identified, and (c)

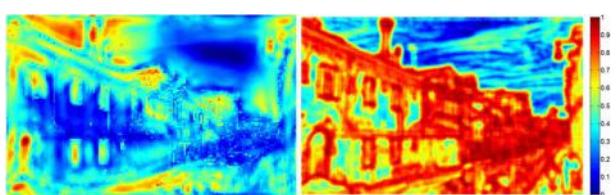


FIGURE 2 | Examples of corresponding saliency and entropy maps (left and right images, respectively) used to simulate gaze patterns in the artificial observer groups (compare to original image in Figure 1). The color legend on the right illustrates the range of possible values for each map.

the saliency value at each of these locations was converted to a probability using the softmax function:

$$\text{Probability of selection} = e^{s_i/\tau} / \sum_{i=1}^{100} e^{s_i/\tau} \quad (1)$$

where s is the given saliency value, and τ is the temperature parameter (fixed at 1). One of these 100 locations on the fixation map was then chosen stochastically, as a function of the corresponding probability values.

This process of updating the inhibition and fixation maps and selecting a new fixation point continued until 13 fixations were performed. The gaze data for 10 artificial observers from the saliency group were then simulated by sweeping through the set of 16 images, once per each observer, and then repeating the process 10 times. It is important to note that repetitions of the simulation process over the same image resulted in distinct gaze patterns, due not only to randomization of the initial fixation, but also to stochasticity in the procedure for selecting subsequent fixations.

Entropy model. The entropy model simulated an artificial observer whose gaze pattern is determined by image “information,” that is, by the presence of structured or organized visual patterns within the image (e.g., Raj et al., 2005; Lin et al., 2010). As a proxy for information, image entropy was estimated for each image. In particular, image entropy reflects the computational cost of compressing an image, based on the frequency of repeated pixel values. The function used for computing image entropy was:

$$\text{Image entropy} = - \sum_{i=1}^{256} p_i * \log_2(p_i) \quad (2)$$

where the original image is converted to grayscale, pixel values are sorted over 256 bins, and p represents the proportion of pixels in each bin.

1. **Entropy maps.** Comparable to the saliency maps, the entropy maps were produced by first downscaling the original images to 480×300 and then converting them to grayscale. Note that the image entropy function produces a single scalar value over the entire image. Thus, the entropy map was produced by sweeping an 11×11 -pixel window over the grayscale image, and replacing the pixel value at the center of the window with the corresponding entropy value for that 11×11 square. **Figure 2** illustrates the entropy map (right image) for one of the outdoor scenes (compare with the original image in **Figure 1**).
2. **Simulated gaze data.** Once the entropy maps were computed for the set of 16 test images, they were then downsampled a second time and normalized, using the same process as described above for the saliency maps. Finally, gaze data for 10 simulated observers were generated, also using the same procedure as described above.

Random model. The random model was designed as a control condition, to simulate the gaze pattern of an observer who explored the test images by following a policy in which all locations are equally-likely to be selected. Thus, no maps were produced for this group. Instead, 2080 x- and y-locations were chosen at random (i.e., 13 fixations \times 16 images \times 10 observers).

Descriptive statistics. We briefly compare here the gaze data produced by each of the five observer groups. In all cases, note that because the random group provides a baseline estimate of performance at chance level, the results from this group are plotted in **Figure 3** as dotted lines (rather than as bars). **Figure 3A** presents the results of projecting each observer group's fixations onto the saliency and entropy maps, respectively, and then computing the average saliency (blue bars) and entropy values (red bars) for the corresponding fixation locations. This analysis provides a measure of the relative influence of saliency vs. entropy for each group's scan patterns. In particular, higher mean values reflect a tendency to orient toward regions in the image with higher levels of saliency and/or entropy, respectively (recall that the values on each map were normalized between 0 and 1). Note that the upper dashed line in **Figure 3A** represents the mean normalized entropy produced by the random observer group, while the lower dashed line represents mean normalized saliency for the same group.

There are three important results. First, as expected, the saliency and entropy observer groups produce near-maximal values (i.e., 90%) for their respective maps. Second, for both infants and adults, the gaze patterns resulted in higher mean levels of entropy than salience. Third, even for the random group, the same pattern was also true. As **Figure 2** suggests, this may be due to differences in how saliency and entropy are distributed over each image—that is, saliency was sparsely distributed while entropy was relatively broadly distributed.

In addition, **Figures 3B–D** present the results of three kinematic measures. First, **Figure 3B** plots the mean dispersion of

fixations for each group. Dispersion was computed by first calculating the centroid of the fixations (i.e., the mean fixation location) within each trial, and then calculating the mean distance of the fixations within that trial from the centroid. As **Figure 3B** indicates, infants tended to have the least-disperse gaze patterns, followed by adults. Interestingly, the dispersion of fixations produced in the saliency observer group was nearly the same as the random observer group.

Next, **Figure 3C** presents the mean gaze shift distance for each group. This distance was calculated by computing how far the fixation point traveled (in pixels) from each fixation to the next. Like the previous result, infants produced the shortest gaze shift distance, again followed by adults. Similarly, the saliency observer group produced gaze shift distances similar to the random observer group, while the entropy observer group had gaze shift distances that fell midway between the real and artificial observers.

Finally, **Figure 3D** presents the mean revisit rate for each observer group. Revisit rate was estimated by first creating a null frequency map (a 480×300 matrix with all locations initialized to zero). Next, for each fixation, the values within a 41×41 square (centered at the fixation location) on the frequency map were incremented by 1. This process was repeated for all of the fixations within a trial, and the frequency map was then divided by the number of fixations. For each trial, the maximum value from this map was recorded, reflecting the location in the image that was *most frequently* visited (as estimated by the 41×41 fixation window). The maximum value was then averaged across trials and observers within each group, providing a metric for the peak proportion of fixations that a particular location in each image was visited, on average. As **Figure 3D** illustrates, a key finding from this analysis is that infants have the highest revisit rate (nearly 50%), while all three of the artificial observer groups have the lowest rates.

COG IMAGE SAMPLES

To maintain tractability of the training set for the SRNs, we randomly selected 20 trials from each group of observers. Selection was subject to several constraints, including: (1) within a group, each observer contributed 2 trials (i.e., gaze data for 2 images), and (2) selection of the corresponding images was counterbalanced both within observer groups and across the 16 images (each image was selected as equally-often as possible across groups). Once the specific trials/images were selected for each group, the gaze data (i.e., sequences of fixation points) were then used to generate the COG training stimuli.

Specifically, for a given observer and trial, a 41×41 grayscale image—centered at the first fixation point—was sampled from the corresponding test image. The dimensions of the COG sample were derived from the display size and viewing distance of the live observers, and correspond to a visual angle of 1.6° , which falls within the estimated range of the angle subtended by the human fovea (Goldstein, 2010). This sampling process continued for the second fixation point, and so on, until the number of fixations for that observer and trial was reached. The process for obtaining the COG samples for a single trial was then repeated through each of the five observer groups, resulting in 20 trials of COG samples per

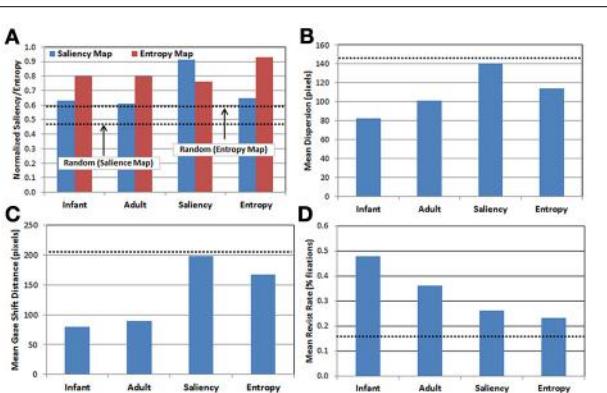
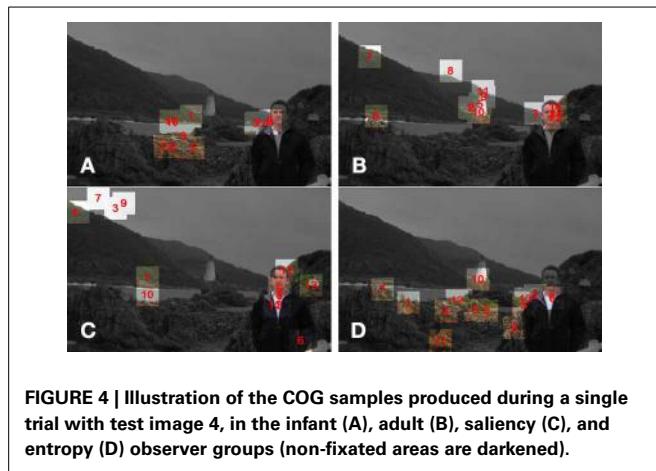


FIGURE 3 | Comparison of gaze patterns across the 5 observer groups (see text for details). (A) Mean map values calculated by projecting each group's gaze points on to the saliency (blue) and entropy (red) maps, respectively; (B) mean dispersion (spread) of fixations; (C) mean gaze shift distance; and (D) mean proportion of revisits. Dashed lines represent performance of the random observer group.



group (with an average of 13 samples per trial, or approximately 260 samples per group).

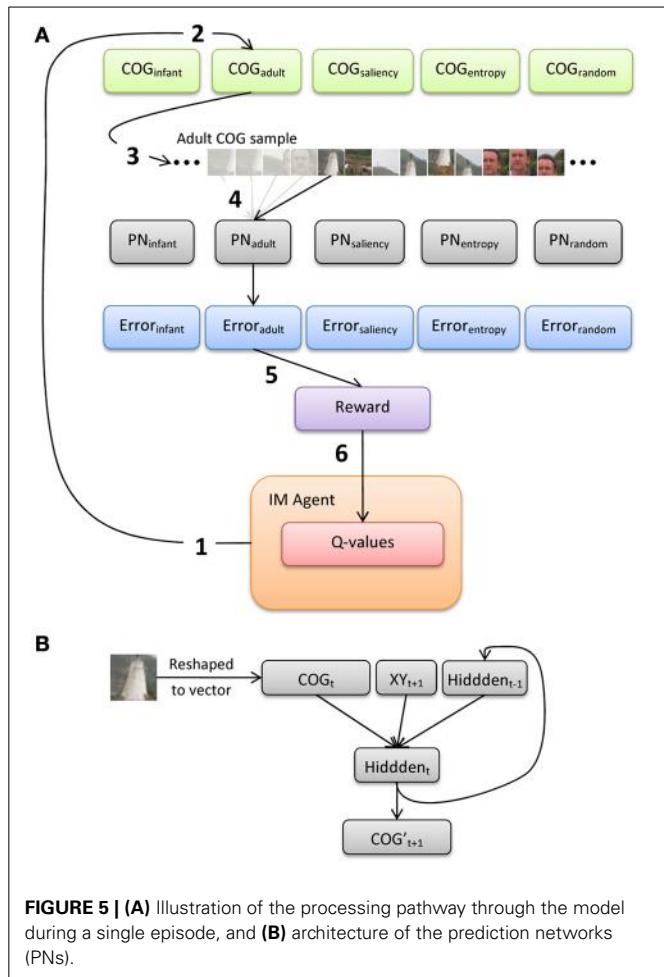
To help illustrate how a typical set of COG samples appears in relation to its corresponding test image, **Figure 4** presents the samples produced during a single trial (with test image 4), in the infant, adult, saliency, and entropy observer groups, superimposed on to the respective test image. Consistent with **Figure 3B**, note that the infant's fixations tend to fall into two spatial clusters, while the adult's fixations are more disperse.

MODEL ARCHITECTURE AND LEARNING ALGORITHMS

Figure 5 illustrates an overview of the model architecture, which implements a conventional reinforcement-learning model layered over a bank of recurrent neural networks. We first provide here a general description of the six major processing steps in the model, and present below a more-detailed description of the PNs and the intrinsically-motivated artificial agent (IM agent).

The IM agent learns over a series of discrete episodes. At the start of each episode (**Figure 5A**, step 1), the IM agent first selects one of the five observer groups. This choice is intended to represent an analog for presenting an image to an observer, who then explores the image by choosing from a set of distinct gaze or scanning "strategies" (alternatively, these strategies could be described as learning goals, behavior or action patterns, etc.). In particular, the IM agent has no direct knowledge of how each strategy is designed or how it operates. Rather, the IM agent bases its decision simply on the current set of Q-values for the set of five choices, which each estimate the long-term sum of rewards expected to result from selecting the corresponding choice. Once one of the gaze-pattern strategies (i.e., observer groups) is selected, the COG samples from the corresponding group of observers are retrieved. For example, in **Figure 5A**, the IM agent selects the adult observer group (step 2).

At the next processing step, the 20 sets of COG samples (from the selected observer group) are then presented to the corresponding SRN (step 3; note that only 1 of the 20 sets is illustrated here). In particular, we implement a bank of five SRNs, each of which is devoted to a single observer group, in order (a) to maintain learnability estimates of all five groups in parallel, and (b) to avoid the risk of catastrophic interference by training a single network on the COG samples from all five groups. We refer to



the SRNs as PNs, as they are explicitly trained to reproduce the series of 41×41 samples, one at a time. In the case of **Figure 5**, one of the 20 COG sample sets is selected at random from the adult observer group, and the first sample from this set is presented to PN_{adult}. The output of the network is its "prediction" of the second sample (properly speaking, since training is offline, i.e., after the samples were collected, the PN learns to *reproduce* a sequence that is iteratively presented). After each output, a training signal is computed using backpropagation-of-error and used to adjust the PN's connection weights. This continues until all of the COG samples in the observer group have been presented to the PN (step 4).

At step 5, the average prediction error for the previous training sweep is computed, and then transformed into a scalar reward value. As we highlight below, we investigate two reward functions: reward based on the magnitude of error (i.e., reward is inversely related to error), and reward based on learning progress (i.e., reduction in error over two consecutive sweeps through the COG samples in an observer group). During the final processing step (6), the new reward value is used to update the set of Q-values, and the IM agent makes its next selection.

PREDICTION NETWORKS

Each PN is a standard 3-layer Elman network, with recurrent connections from the hidden layer back to the input layer (i.e.,

context units; Elman, 1990). In particular, the PN implements a forward model, in which the current sensory input (plus a planned action) is used to generate a prediction of the next expected input (e.g., Jordan and Rumelhart, 1992). Prior to training the PN, each of the COG samples is converted to grayscale values between 0 and 1. As **Figure 5B** illustrates, the input layer is composed of 2083 units, including a vector of 1681 units that encode the grayscale pixel values of the COG sample, 2 units that encode the (normalized) x- and y-coordinates of the upcoming COG sample, and 400 context units (which copy back the activity of the hidden layer from the previous time step). There are 400 units in the hidden layer (i.e., roughly 75% compression of the input) and 1681 output units.

All connections in the PN are initialized with random values between 0 and 1, which are then divided by the number of incoming units (i.e., fan-in). For each simulation run, the same PN is cloned five times, so that all five PNs begin with the same set of initial connection weights. As noted above, each PN is presented with only the COG samples from its corresponding observer group. Once an observer group is selected by the IM agent, the 20 COG sample sets are then presented to the appropriate PN in random order. Recall that each set of COG samples represents the gaze data from a single observer and a single trial. In order to remove the influence of previous trials on the context layer activation, the units in the context layer of the PN are initialized to 0.5 at the start of each trial. A single training epoch is defined as a sweep through all 20 trials.

Prediction error is measured as the root mean-squared error (RMSE), computed over the difference between each predicted and observed next COG sample, and then averaged over the entire trial. Mean trial errors are then averaged together over the 20 trials; this value represents the mean prediction error for the IM agent's current episode, and is used to compute the reward signal.

IM AGENT

The IM agent simulates a naïve, active observer that is reinforced for visually exploring its environment. As **Figure 5** illustrates, the IM agent is provided with the opportunity to select among five predefined sets of visual samples and a corresponding PN, each of which represents (ostensibly) a unique scanning experience and learning episode over the set of 16 test images. After each selection, the IM agent receives a reward signal as feedback that is proportional—not to the content or the quality of the chosen gaze samples *per se*—but rather, to the relative success of the chosen PN in predicting the resulting sequence of COG samples. In other words, the IM agent is rewarded for choosing the set of COG samples (i.e., a pattern of visual exploration) that is learned optimally.

In principle, defining an *exploration reward* on the basis of *learnability* runs the risk of generating an unintended outcome. For example, one way to maximize the performance of the PN is to hold the fixation point constant, that is, to continue looking at the same location. Such a strategy, however, also provides limited visual information (i.e., it maximizes prediction but minimizes exploration). At the other extreme, a completely random gaze sequence may be highly informative, but difficult, if not impossible to predict. Given the putative goal of visual exploration, therefore, a reasonable trade-off is to select a gaze sequence that is

both informative *and* predictable (i.e., varied but also systematically structured). We therefore, note here that linking the reward function to prediction learning captures an important dimension of visual exploration, but that other facets such as novelty are also likely to play a role (for a comprehensive discussion of knowledge-based vs. competence-based approaches to intrinsic motivation, see Oudeyer and Kaplan, 2007, and Baldassarre and Mirolli, 2013).

Because the actions selected by the IM agent are influenced by the performance of the PNs, there are effectively two timescales: an “inner loop,” which is defined as presenting the selected PN with the COG samples from a single trial, and the “outer loop,” which is a single episode and is defined as the IM agent's selection of an observer group, a training epoch of the corresponding PN, the generation of an intrinsic reward signal, and the updating of the IM agent's Q-values (as illustrated in **Figure 5**). For both Simulations 1 and 2, therefore, a single simulation run included 500 iterations of the outer loop (i.e., episodes). In addition, recall that during each iteration of the outer loop, there were 20 iterations of the inner loop for the selected PN.

As we highlight below, the objective or reward function that we implemented was varied across simulations. In Simulation 1, the reward was defined as:

$$r_t = 1 - \text{Error}_t \quad (3)$$

where r_t is the reward received for the t th iteration of the outer loop, and Error_t is the mean error produced by the PN selected during iteration t . This function therefore, rewards the IM agent for selecting the observer group with the lowest prediction errors (compare to “predictive novelty,” i.e., Equation 9 in Oudeyer and Kaplan, 2007). In contrast, during Simulation 2 the reward function was defined as the percent change in prediction error over two consecutive iterations of the inner loop:

$$r_t = (\text{Error}_{t-1} - \text{Error}_t) / \text{Error}_{t-1}$$

where Error_t is defined as in Equation (3), and Error_{t-1} represents the corresponding mean error from the previous iteration. Note that in this case, each time a PN was selected, it was trained for two consecutive epochs before the IM agent received a reward.

Two steps were implemented to ensure that the IM agent sufficiently explored each of the five observer groups. First, at the start of each simulation run, the IM agent's Q-values were initialized optimistically, that is, they were set to initial values higher than were expected to occur during learning. Second, the Softmax function [see Equation (1)] was used for action selection, which provided an additional source of stochasticity and variability into the IM agent's choice of observer group.

After selecting an observer group and receiving a reward for the selection, the IM Agent's Q-value for that group was updated. The update rule implemented was:

$$Q_t = Q_{t-1} + \alpha(r_t - Q_{t-1}) \quad (4)$$

where Q_{t-1} is the Q-value for the selected observer group before the most recent iteration of the inner loop, and Q_t is the new, updated value after the iteration. Finally, α represents the learning rate, which was fixed for each simulation.

SIMULATION 1

In Simulation 1, the IM agent vicariously explored the 16 test images by repeatedly selecting from a set of COG samples, each of which captured the process of scanning the images in either real or simulated real time. After each selection, the IM agent then received a reward which represented the relative ease or difficulty of sequentially predicting the selected gaze samples. In particular, the IM agent received a larger reward when it picked a set of COG samples that were “easily” learned (i.e., that resulted in comparatively lower prediction errors), while the scalar reward was lower when the COG samples (and the corresponding PN) produced higher prediction errors. Our primary prediction was that, given the assumption that infants are mastering the skill of visual exploration, the COG samples produced by the 9-month-olds would be the most predictable, and therefore, the IM agent would prefer samples produced by the 9-month-olds over those from the other four observer groups.

METHOD

Ten simulation runs were conducted. At the start of each run, the five PNs were initialized as described above. In addition, the set of Q-values for the five corresponding actions was uniformly initialized to 1. During Simulation 1, the temperature parameter τ used in the Softmax function for action selection was 0.01. Finally, the learning rate value α used for updating the Q-values (Equation 5) was 0.1. Each simulation run was composed of 500 episodes, during each of which the IM agent chose a set of COG samples, the corresponding PN was trained on the selected set of samples for one epoch, and the IM agent then received a reward and the respective Q-value was updated.

RESULTS

For the purpose of analysis, the results over the 10 simulation runs were averaged together. We focus here on three questions. First, during learning, does the IM agent develop a preference for any of the five observer groups? Second, how does the IM agent distribute its selections over the five groups? Finally, how well do the five PNs collectively perform over the 500 episodes?

We addressed the first question by transforming the Q-values at the end of each episode into standardized “preference” values, which are simply the probabilities assigned to the choices by the Softmax function. **Figure 6A** presents the mean preferences for the five observer groups as a function of episode, averaged across 10 simulation runs. Mean preferences were analyzed statistically by dividing the 500 training episodes into 10 blocks, each 50 episodes long. We then conducted a two-factor mixed-model ANOVA for each of the blocks, with observer group (infant, adult, saliency, entropy, and random) as the between-subjects factor, and episode as the within-subjects factor. We report here the results of the planned paired-comparison tests for the five observer groups, focusing specifically on whether the group (or groups) with the highest preference values differed significantly from the remaining observer groups. Note that the top legend in **Figure 6A** illustrates the outcome of these comparisons for each of the 50-episode blocks, by indicating the group/groups with the highest preference value and the significance level of the planned comparison (I = infant, A = adult, S = saliency, E = entropy, R = random).

There were three major findings. First, for approximately the first 50 episodes, preference values varied considerably, resulting in no significant differences between the five observer groups. Second, a preference for the COG samples from the infant observer group emerged between episodes 50 and 100, while the values for the other four groups continued to decline. Third, and confirming our prediction, this pattern continued and strengthened between episodes 100 and 500.

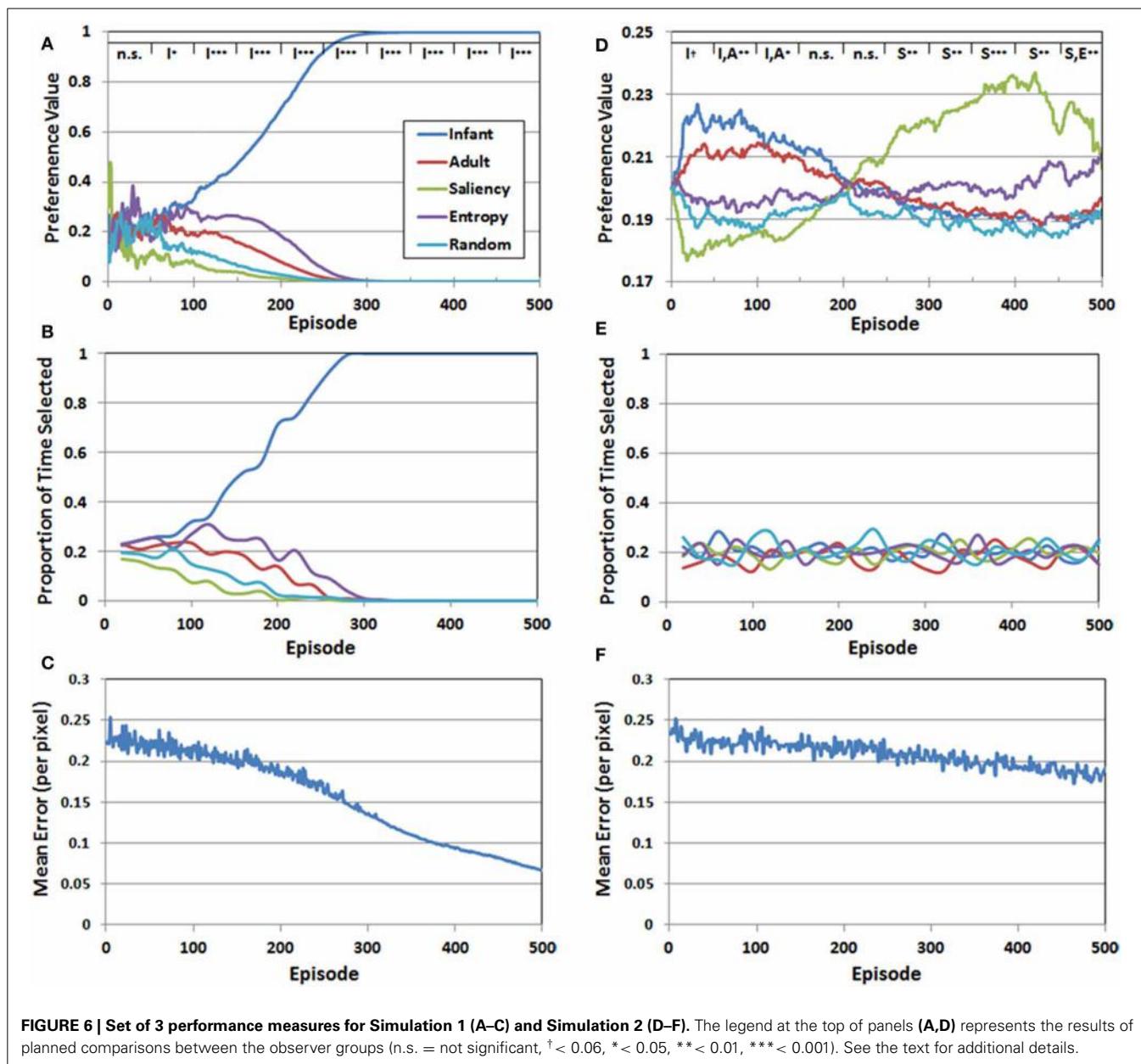
Figure 6B presents the proportion of time that each of the five observer groups was selected over the 500 episodes. Recall that because a stochastic decision rule was used to select the groups, the actual frequency of selection may not necessarily align with the corresponding preference values. However, as **Figure 6B** illustrates, there was a close match between the IM agent’s preference values, and the resulting selection pattern. In particular, during the last 200 episodes, effectively all of the training time was directed toward the infant observer group’s PN.

Finally, **Figure 6C** presents the RMSE—pooled over the five PNs—as a function of episode. At the start of training, the RMSE was approximately 0.25 per pixel. Fluctuations in the error level, between episodes 1 and 300, reflected the fact that the IM agent continued to explore the observer groups throughout this period. However, as the infant observer group became the sole preferred choice, the IM agent focused on the COG samples from this group and the error rate declined more consistently. By 500 episodes, the RMSE had fallen below 0.07. Thus, **Figure 6C** suggests that all of the PNs improved during training, but the infant group’s PN eventually received the majority of training time and accordingly benefited.

SIMULATION 2

While Simulation 1 confirmed our prediction that the IM agent would prefer the infant observer group’s COG samples, it is also important to note that the particular reward function used potentially suffers from a “snowball” bias. In other words, because the reward function favored low prediction errors, the group with the lowest errors at the start of training would have an advantage over the other four groups. In addition, a bias toward providing this group with additional training time would then continue to improve the predictions of their PN, thereby lowering prediction errors further and increasing the advantage of that group. Such a bias would also reduce exploration of the competing groups, and consequently, leave them with higher errors.

To address this issue, we investigated an alternative reward function, which favored learning progress, that is, a reduction in the RMSE over two consecutive episodes. As Equation 4 highlights, the reward function in Simulation 2 was scaled by the RMSE of the first episode of each pair, which effectively produced a reward value equal to the percent change in the RMSE. Interestingly, this solves one problem while creating a new challenge for the model: in particular, by linking reward to changes in performance of the PNs, the IM agent’s learning task becomes non-stationary. Specifically, by selecting the “best” (i.e., most-improving) observer group for training, learning in that group should eventually level off, and thus, the IM agent’s long-term estimates of the group’s Q-value should systematically drift downward over time. Fortunately, there is also a hidden advantage to this approach, namely, that the IM agent should therefore,



switch its preference from the COG samples of one observer group to another, as improvement in the leading group slows. As we highlight in the discussion, such a switching pattern has the potential to be interpreted as a developmental pattern, in which the simulated observer shifts from one visual-exploration strategy to another.

Recall that our prediction for Simulation 2 was that, like Simulation 1, the COG samples from the infant observer group would be preferred first, and that the model would then shift its preference to the samples from the adult observer group.

METHOD

The same procedures as Simulation 1 were followed in Simulation 2. However, given an expected decline in the absolute magnitude of the reward (relative to Simulation 1), the Softmax parameter τ was increased to 0.1, the initial Q -values were lowered to

0.01, and the learning rate value α used for updating the Q -values was lowered to 0.05. In addition, as noted above, the IM agent selected an observer group on every odd-numbered episode, and then received a reward value after the subsequent even-numbered episode. Training of the PNs continued, as in Simulation 1, for all episodes.

RESULTS

Figure 6D presents the mean preference values for the five observer groups in Simulation 2, as a function of episode number. These values were analyzed following the same analytical strategy described in Simulation 1. A key finding from the analysis is that the range of preference values was considerably narrower than the pattern observed in Simulation 1. In addition, although we predicted that the COG samples from the infant observer group would have the highest initial preference values,

this preference was not as robust as we anticipated. In particular, there was a marginally-significant preference for the infant observer group ($p < 0.06$) between episodes 1 and 50. Between episodes 50 and 100, there was no longer a significant difference between the infant and adult observers, though the two real observer groups had significantly higher preference values than the artificial observer groups ($p < 0.01$). This pattern maintained through episode 150. For the next 100 episodes (150–250) there was no significant difference between the five groups. Between episode 250 and 300, the leading preference shifted to the saliency observer group. This pattern persisted through the remaining episodes, although as **Figure 6D** illustrates, the preference values for the entropy observer group increased toward the end of training.

In contrast to Simulation 1, in which a clear preference for one of the observer groups was matched by a tendency for the corresponding group to also be selected consistently by the IM agent, there was a comparatively narrower preference pattern in Simulation 2, and as **Figure 6E** illustrates, also lack of a clear selection pattern. Indeed, the proportion of times each group was selected in Simulation 2 continued to fluctuate throughout the entire simulation.

Finally, **Figure 6F** presents the RMSE (pooled over observer groups) generated by the PNs over 500 episodes. In contrast to **Figure 6C**, the error rate declined more slowly in Simulation 2. There are several factors that may have contributed to this pattern. First, as noted above, the IM agent continued to explore until the end of Simulation 2, while in Simulation 1, exploratory selection of the sub-optimal observer groups ended on average by the 300th episode. Another contributing factor is that the relative differences in the five Q-values were smaller in Simulation 2, which also increased the chances of exploratory selections. Indeed, as we expected, there was no sustained “winner,” but rather, a series of shifts from one observer group to another.

However, it should be noted the second observer group that became preferred by the IM agent (i.e., after episode 250) was *not* the adult observer group, as we predicted. Instead, as **Figure 6D** illustrates, it was instead the saliency observer group. This result raises an important and interesting property of the reward function used in Simulation 2. In particular, note that the saliency observer group is the *least* preferred in Simulation 1, which is ostensibly due to having the largest initial prediction errors. Nevertheless, these initially high prediction errors may have helped to make the saliency observer group stand out in Simulation 2, as the COG samples from this group presumably provided the second-best opportunity for the IM agent to optimize its learning progress.

GENERAL DISCUSSION

We provided an artificial agent with the opportunity to select among five sets of visual-exploration patterns, and then reinforced the agent for selecting COG samples that were either the easiest to learn (Simulation 1), or afforded the largest improvements in learning (Simulation 2), as estimated by a prediction-learning model. The agent was intrinsically-motivated, in the sense that it was not solving an explicit task—such as locating an object in a visual scene or comparing two images—but rather, it was rewarded for how well it learned (or

more accurately, how well it selected a set of training images together with an artificial neural network that learned the set).

The pattern of findings from two simulation studies confirmed the first of three predictions, and partially confirmed the second. First, in Simulation 1—where the reward function was based on minimizing prediction errors—we found that the IM agent showed a consistent preference for learning from the COG image samples that were produced by human infants, rather than those produced by human adults, or those from three groups of artificial observers. Second, in Simulation 2 we predicted that infants’ COG image samples would initially be preferred, and that the IM agent would then switch its preference to the adult observer group. While the first half of the prediction was confirmed, there were two qualifications: (a) the initial preference for the infant observer group was only marginally significant, and (b) this preference soon gave way to a collective preference for both the infant and adult COG image samples—that is, a preference for the real observer groups over the artificial observer groups. We also did not observe a clear switch to the adult observer group. Instead and contrary to our third prediction, the second preference “wave” in Simulation 2 was for the saliency observer group. While the data collected in the present study may not provide a comprehensive explanation for this result, we note below that our previous work highlights the important role of image salience, and may ultimately provide some insight into the pattern of findings in Simulation 2.

There are a number of implications for understanding development, as well as important questions, which are raised by these findings. First, our results suggest that if (1) prediction-learning and future-oriented actions play a central role in early visual development, and (2) infants are intrinsically-motivated to fine-tune and improve their ability to predict or forecast upcoming events, then the gaze patterns produced by 9-month-olds are well-suited to achieving both of those goals, compared to the gaze patterns of adults or the artificial observers that we generated. However, this finding also raises the question: what are the features of 9-month-olds’ gaze patterns that make their COG samples easier to learn than those of other observers?

The kinematic analyses presented in **Figure 3** suggest that how infants distribute their gaze over space may provide an important clue to answering this question. One possibility is that because 9-month-olds tend to have less-disperse gaze patterns than adults, and to shift their gaze a shorter distance, the resulting COG samples they produce tend to be more homogenous, and therefore, easier to learn. Alternatively, it may be the case that infants have the *a priori* goal of generating easily-learnable gaze patterns, and as a result, they therefore, tend to produce more compact scanpaths, with shorter gaze shifts between fixations. An essential step toward addressing this “chicken-and-egg” question is to collect gaze samples from a wider range of infants (e.g., 3- and 6-month-olds) and to evaluate the model when those additional COG samples are included. Another approach is to pit gaze-travel distance against local/global similarity, by using carefully-designed test images, in which there is high variability at the local level, with sets of highly-similar regions that are spaced relatively far apart.

A second issue suggested by our findings is what the developmental pattern will look like when the gaze data from younger

infants are included. For example, should the agent prefer 3-month-olds' COG samples over those from 9-month-olds? In principle, with data from infants between birth and 12 months, our intuition is to expect an inverted U-shaped developmental pattern, in which gaze data from very young infants is poorly-controlled and therefore, highly unpredictable. We would then expect maximally-predictable COG samples between 3 and 4 months, and then an increasing trend afterwards of gradually less and less predictable gaze patterns. Fortunately, this is an empirical question that can be tested without any major modifications to our model.

Finally, a third question is whether the pattern of results—in particular, the shift that we observed during Simulation 2—can be interpreted as implying a *developmental pattern*. This is a difficult question to answer, as the timescale of the simulation reflects learning in an artificial agent, and does not map directly onto the infant-developmental timeline. Nevertheless, we might “read off” the results from Simulation 2 as suggesting that an initial strategy for visual exploration during infancy is to first focus on producing relatively dense clusters of fixations (i.e., like those produced by the two real-observer groups), which then shift toward becoming more widely distributed, and in particular, increasingly sensitive to the presence of salient regions in the visual scene. While this issue remains an open question, our prior work demonstrates that image saliency is an important factor that successfully accounts for infants' performance on a number of perceptual tasks (e.g., Schlesinger et al., 2007, 2011, 2012).

There are also a number of ways that our current approach can be improved. First, it is important to note that the PNs were trained offline—that is, the networks were trained to predict gaze sequences that had already been collected or generated. A disadvantage of this method is that any changes that occur in the agent cannot be propagated back to the observer groups. In other words, while the agent influences the amount of training time that each PN receives, it cannot influence how the COG samples are produced. An alternative and perhaps more-informative design would be for the choices of the agent to have an impact on the COG sampling process itself. Indeed, such a mechanism could be designed so that the production of eye movements in the artificial model is linked to the choices of the agent. However, there is no obvious way in which a similar connection could also be made between the agent and a live observer.

A second limitation of our model is that five different PNs were employed, which might be interpreted to suggest that infants' generate multiple sets of parallel predictors during visual exploration and then sample among them. While we remain agnostic

to the specific cognitive structures or architectures exploited by human infants during visual exploration, a more elegant solution on the computational side would be to employ a single, unified predictor that learns over a range of sampling strategies (e.g., Schmidhuber, 2010).

Finally, a third issue concerns the models of the artificial observers, and in particular, the procedure used to transform the saliency and entropy maps into sequences of simulated eye movements. A key difference between the artificial and real observers is that the artificial observers tended to produce more disperse fixations, and return to previously-fixated locations less often than the human infants and adults. This issue can be addressed by imposing a theoretical energy or metabolic “cost” to the simulated eye movements, which is proportional to the size of the saccade. In addition, we can also continue to tune and improve the IOR mechanism, perhaps by modifying the decay rate, so that inhibition for previously-fixated locations decreases more rapidly. Another promising approach is to “yoke” the simulated gaze data to the actual moment-to-moment eye movements produced by real observers, so that kinematic measures such as fixation duration or saccade size are matched across the real and artificial data sets.

We conclude by noting that our work thus far takes advantage of machine-learning methods—in particular, the set of learning algorithms and architectures used to study intrinsic motivation in natural and artificial systems—as a means toward the goal of understanding visual development in human infants. Nevertheless, it is important to stress that the influence also runs in the other direction, that is, what lessons can be taken from our approach that might prove useful to the design of robots and artificial agents? One interesting insight is that our findings are consistent with the idea of “starting small” (e.g., Elman, 1993; Schlesinger et al., 2000); in other words, infants' gaze patterns may provide an advantageous starting point for learning in a naïve agent, relative to more-experienced observers such as adults. As we continue to extend and elaborate our model, in particular with data from younger infants, we anticipate that other important lessons for designing and developing artificial agents will continue to emerge.

ACKNOWLEDGMENTS

This project received support from the National Institute of General Medical Sciences, National Institutes of Health (P20GM103645) and the James S. McDonnell Foundation Scholar Award to Dima Amso. Additional support was provided to Matthew Schlesinger by the SIUC Cope Fund.

REFERENCES

- Amso, D., Haas, S., Tenenbaum, E., Markant, J., and Sheinkopf, S. J. (2013). Bottom-up attention orienting in young children with autism. *J. Autism Dev. Disord.* 43, 1–10. doi: 10.1007/s10803-013-1925-5
- Amso, D., and Johnson, S. P. (2005). Selection and inhibition in infancy: evidence from the spatial negative priming paradigm. *Cognition* 95, B27–B36. doi: 10.1016/j.cognition.2004.08.006
- Aslin, R. (2009). How infants view natural scenes gathered from a head-mounted camera. *Optom. Vis. Sci.* 86, 561. doi: 10.1097/OPX.0b013e3181a76e96
- Baldassarre, G. (2011). “What are intrinsic motivations? A biological perspective,” in *Proceedings of the International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob-2011)*, eds A. Cangelosi, J. Triesch, I. Fasel, K. Rohlfing, F. Nori, P.-Y. Oudeyer, M. Schlesinger, and Y. Nagai (New York, NY: IEEE).
- Baldassarre, G., and Mirolli, M. (2013). “Intrinsically motivated learning systems: an overview,” in *Intrinsically Motivated Learning in Natural and Artificial Systems*, eds G. Baldassarre and M. Mirolli (Berlin: Springer-Verlag), 1–14. doi: 10.1007/978-3-642-32375-1
- Bronson, G. (1982). *The Scanning Patterns of Human Infants: Implications for Visual Learning*. Norwood, NJ: Ablex.
- Bronson, G. (1991). Infant differences in rate of visual encoding. *Child Dev.* 62, 44–54. doi: 10.2307/1130703

- Bushnell, I. W. R., Sai, F., and Mullin, J. T. (1989). Neonatal recognition of the mother's face. *Br. J. Dev. Psychol.* 7, 3–15. doi: 10.1111/j.2044-835X.1989.tb00784.x
- Dragoi, V., and Sur, M. (2006). Image structure at the center of gaze during free viewing. *J. Cogn. Neurosci.* 18, 737–748. doi: 10.1162/jocn.2006.18.5.737
- Elman, J. L. (1990). Finding structure in time. *Cogn. Sci.* 14, 179–211. doi: 10.1207/s15516709cog1402_1
- Elman, J. L. (1993). Learning and development in neural networks: the importance of starting small. *Cognition* 48, 71–99. doi: 10.1016/0010-0277(93)90058-4
- Fantz, R. L. (1956). A method for studying early visual development. *Percept. Mot. Skills* 6, 13–15. doi: 10.2466/pms.1956.6.g.13
- Field, D. J. (1994). What is the goal of sensory coding? *Neural Comput.* 6, 559–601. doi: 10.1162/neco.1994.6.4.559
- Frank, M. C., Vul, E., and Johnson, S. P. (2009). Development of infants' attention to faces during the first year. *Cognition* 110, 160–170. doi: 10.1016/j.cognition.2008.11.010
- Frank, M. C., Vul, E., and Saxe, R. (2012). Measuring the development of social attention using free-viewing. *Infancy* 17, 355–375. doi: 10.1111/j.1532-7078.2011.00086.x
- Goldstein, B. (2010). *Sensation and Perception*. Belmont, CA: Wadsworth.
- Haith, M. M. (1980). *Rules that Babies look by: The Organization of Newborn Visual Activity*. New Jersey: Erlbaum.
- Haith, M. M. (1994). *The Development of Future-Oriented Processes*. Chicago: University of Chicago Press.
- Haith, M. M., Hazan, C., and Goodman, G. S. (1988). Expectation and anticipation of dynamic visual events by 3.5-month-old babies. *Child Dev.* 59, 467–479. doi: 10.2307/1130325
- Hayhoe, M., and Ballard, D. (2005). Eye movements in natural behavior. *Trends Cogn. Sci.* 9, 188–194. doi: 10.1016/j.tics.2005.02.009
- Itti, L., and Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Res.* 40, 1489–1506. doi: 10.1016/S0042-6989(99)00163-7
- Itti, L., Koch, C., and Niebur, E. (1998). A model of saliency-based visual-attention for rapid scene analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* 20, 1254–1259. doi: 10.1109/34.730558
- Johnson, S. P., Amso, D., and Slemmer, J. A. (2003). Development of object concepts in infancy: evidence for early learning in an eye-tracking paradigm. *Proc. Natl. Acad. Sci. U.S.A.* 100, 10568–10573. doi: 10.1073/pnas.1630655100
- Johnson, S. P., Slemmer, J. A., and Amso, D. (2004). Where infants look determines how they see: eye movements and object perception performance in 3-month-olds. *Infancy* 6, 185–201. doi: 10.1207/s15327078in0602_3
- Jordan, M. J., and Rumelhart, D. E. (1992). Forward models: supervised learning with a distal teacher. *Cogn. Sci.* 16, 307–354. doi: 10.1207/s15516709cog1603_1
- Kenward, B. (2010). 10-month-olds visually anticipate an outcome contingent on their own action. *Infancy* 15, 337–361. doi: 10.1111/j.1532-7078.2009.00018.x
- Klatzky, R. L., and Lederman, S. (1990). "Intelligent exploration by the human hand," in *Dextrous Robot Hands*, eds S.T. Venkataraman and T. Iberall (New York, NY: Springer), 66–81. doi: 10.1007/978-1-4613-8974-3_4
- Lin, Y., Fang, B., and Tang, Y. (2010). "A computational model for saliency maps by using local entropy," in *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence* (Atlanta, GA), 967–973.
- Maurer, D., and Barrera, M. (1981). Infants' perception of natural and distorted arrangements of a schematic face. *Child Dev.* 47, 523–527. doi: 10.2307/1128813
- Mohammed, R. A. A., Mohammed, S. A., and Schwabe, L. (2012). BatGaze: a new tool to measure depth features at the center of gaze during free viewing. *Brain Informatics* 7670, 85–96. doi: 10.1007/978-3-642-35139-6_9
- Morton, J., and Johnson, M. H. (1991). Conspec and conlern: a two-process theory of infant face recognition. *Psychol. Rev.* 98, 164–181. doi: 10.1037/0033-295X.98.2.164
- Oudeyer, P.-Y., and Kaplan, F. (2007). What is intrinsic motivation? A typology of computational approaches. *Front. Neurorobot.* 1:6. doi: 10.3389/neuro.12.006.2007
- Perone, S., and Spencer, J. P. (2013). Autonomous visual exploration creates developmental change in familiarity and novelty seeking behaviors. *Front. Psychol.* 4:648. doi: 10.3389/fpsyg.2013.00648
- Raj, R., Geisler, W. S., Frazor, R. A., and Bovik, A. C. (2005). Contrast statistics for foveated visual systems: fixation selection by minimizing contrast entropy. *J. Opt. Soc. Am. A Opt. Image Sci. Vis.* 22, 2039–2049. doi: 10.1364/JOSAA.22.002039
- Saffran, J. R., Aslin, R. N., and Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science* 274, 1926–1928. doi: 10.1126/science.274.5294.1926
- Schlesinger, M. (2013). "Investigating the origins of intrinsic motivation in human infants," in *Intrinsically Motivated Learning in Natural and Artificial Systems*, eds G. Baldassarre and M. Mirolli (Berlin: Springer-Verlag), 367–392. doi: 10.1007/978-3-642-32375-1_14
- Schlesinger, M., Amso, D., and Johnson, S. P. (2007). The neural basis for visual selective attention in young infants: a computational account. *Adapt. Behav.* 15, 135–148. doi: 10.1177/1059712307078661
- Schlesinger, M., Amso, D., and Johnson, S. P. (2011). "Increasing spatial competition enhances visual prediction learning," in *Proceedings of the First Joint IEEE Conference on Development and Learning and on Epigenetic Robotics*, eds A. Cangelosi, J. Triesch, I. Fasel, K. Rohlfing, F. Nori, P.-Y. Oudeyer, M. Schlesinger, and Y. Nagai (New York, NY: IEEE).
- Schlesinger, M., Amso, D., and Johnson, S. P. (2012). Simulating the role of visual selective attention during the development of perceptual completion. *Dev. Sci.* 15, 739–752. doi: 10.1111/j.1467-7687.2012.01177.x
- Schlesinger, M., Parisi, D., and Langer, J. (2000). Learning to reach by constraining the movement search space. *Dev. Sci.* 3, 67–80. doi: 10.1111/1467-7687.00101
- Schmidhuber, J. (2010). Formal theory of creativity, fun, and intrinsic motivation (1990–2010). *IEEE Trans. Auton. Ment. Dev.* 2, 230–247. doi: 10.1109/TAMD.2010.2056368
- Shinoda, H., Hayhoe, M. M., and Shrivastava, A. (2001). What controls attention in natural environments? *Vision Res.* 41, 3535–3545. doi: 10.1016/S0042-6989(01)00199-7
- Slater, A. (2002). Visual perception in the newborn infant: issues and debates. *Intellectica* 34, 57–76. Available online at: http://intellectica.org/SiteArchives/archives/n34/n34_table.htm
- Triesch, J., Jasso, H., and Deak, G. O. (2007). Emergence of mirror neurons in a model of gaze following. *Adapt. Behav.* 15, 149–165. doi: 10.1177/1059712307078654
- von Hofsten, C. (2010). Prospective control: a basic aspect of action development. *Hum. Dev.* 36, 253–270. doi: 10.1159/000278212
- Wang, Q., Bolhuis, J., Rothkopf, C. A., Kolling, T., Knopf, M., and Triesch, J. (2012). Infants in control: rapid anticipation of action outcomes in a gaze-contingent paradigm. *PLoS ONE* 7:e30884. doi: 10.1371/journal.pone.0030884

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 21 June 2013; accepted: 10 October 2013; published online: 31 October 2013.

*Citation: Schlesinger M and Amso D (2013) Image free-viewing as intrinsically-motivated exploration: estimating the learnability of center-of-gaze image samples in infants and adults. *Front. Psychol.* 4:802. doi: 10.3389/fpsyg.2013.00802*

This article was submitted to Cognitive Science, a section of the journal Frontiers in Psychology.

Copyright © 2013 Schlesinger and Amso. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Which is the best intrinsic motivation signal for learning multiple skills?

Vieri G. Santucci^{1,2*}, Gianluca Baldassarre¹ and Marco Mirolli¹

¹ Laboratory of Computational Embodied Neuroscience, Istituto di Scienze e Tecnologie della Cognizione, Consiglio Nazionale delle Ricerche, Roma, Italy

² School of Computing and Mathematics, University of Plymouth, Plymouth, UK

Edited by:

Tom Stafford, University of Sheffield, UK

Reviewed by:

Fabien Hervouet, Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier, France

Yiannis Gatsoulis, University of Ulster, UK

***Correspondence:**

Vieri G. Santucci, Laboratory of Computational Embodied Neuroscience, Istituto di Scienze e Tecnologie della Cognizione, Consiglio Nazionale delle Ricerche, Via San Martino della Battaglia 44, Roma 00185, Italy
e-mail: vieri.santucci@istc.cnr.it

Humans and other biological agents are able to autonomously learn and cache different skills in the absence of any biological pressure or any assigned task. In this respect, Intrinsic Motivations (i.e., motivations not connected to reward-related stimuli) play a cardinal role in animal learning, and can be considered as a fundamental tool for developing more autonomous and more adaptive artificial agents. In this work, we provide an exhaustive analysis of a scarcely investigated problem: which kind of IM reinforcement signal is the most suitable for driving the acquisition of multiple skills in the shortest time? To this purpose we implemented an artificial agent with a hierarchical architecture that allows to learn and cache different skills. We tested the system in a setup with continuous states and actions, in particular, with a kinematic robotic arm that has to learn different reaching tasks. We compare the results of different versions of the system driven by several different intrinsic motivation signals. The results show (a) that intrinsic reinforcements purely based on the knowledge of the system are not appropriate to guide the acquisition of multiple skills, and (b) that the stronger the link between the IM signal and the competence of the system, the better the performance.

Keywords: intrinsic motivations, learning signals, multiple skills, hierarchical architecture, competence acquisition, reinforcement learning, simulated robot

1. INTRODUCTION

The ability to learn and cache multiple skills in order to use them when required is one of the main characteristics of biological agents: forming ample repertoires of actions is important to widen the possibility for an agent to better adapt to different environments and to improve its chance of survival and reproduction.

Moreover, humans and other mammals (e.g., rats and monkeys) explore the environment and learn new skills not only on the basis of reward-related stimuli but also on the basis of novel or unexpected neutral stimuli. The mechanisms related to this kind of learning processes have been studied since the 1950s, first in animal psychology (e.g., Harlow, 1950; White, 1959) then in human psychology (e.g., Berlyne, 1960; Ryan and Deci, 2000), under the heading of “Intrinsic Motivations” (IMs). Recently, researchers have also started to investigate the neural basis of those mechanisms, both through experiments (e.g., Wittmann et al., 2008; Duzel et al., 2010) and computational models (e.g., Kakade and Dayan, 2002; Mirolli et al., 2013), and IMs are nowadays an important field of research (Baldassarre and Mirolli, 2013a).

From a computational point of view, IMs can be considered a useful tool to improve the implementation of more autonomous and more adaptive artificial agents and robots. In particular, IM learning signals can drive the acquisition of different skills without any assigned reward or task. Most of the IM computational models are implemented within the framework of reinforcement learning (Sutton and Barto, 1998) and, following the seminal

works of Schmidhuber (1991a,b), most of them implement IMs as intrinsic reinforcements based on the prediction error (PE), or on the improvement in the prediction error (PEI), of a predictor of future states of the world.

Despite the increasing number of computational researches based on IMs (e.g., Barto et al., 2004; Schembri et al., 2007b; Oudeyer et al., 2007a; Santucci et al., 2010; Baranes and Oudeyer, 2013), it is not yet clear which kind of IM reinforcement signal is the most suitable for driving a system to learn the largest number of skills in the shortest time. To our knowledge, there are only few studies dedicated to this important issue (Lopes and Oudeyer, 2012; Santucci et al., 2012a, 2013b). In our previous works (Santucci et al., 2012a, 2013b), we have shown the importance of coupling the activity of the mechanism generating the IM signal to the competence of the system in performing the different tasks. However, in Santucci et al. (2012a) we limited our analysis to the learning of a single skill in a simple grid-world environment, while in Santucci et al. (2013b), although implementing a hierarchical architecture able to learn multiple tasks within continuous states and actions spaces, we focused only on signals based on PE. Lopes and Oudeyer (2012) deal with a similar problem, i.e., learning n tasks in the best possible way within a limited amount of time. The solution they propose is to allocate each unit of learning time to the task that guarantees the maximum improvement. However, their work tackles the problem in an abstract and disembodied setup and, moreover, they assume that the system has the information on the learning curves of each task.

In this work, we provide an exhaustive analysis of this scarcely investigated problem: which kind of IM reinforcement signal is the most suitable for driving the learning of skills in the shortest time. With this work we also aim to validate our hypothesis on the importance of a close coupling between the IM learning signal and the actual competence of the system in learning the different tasks. To this purpose, we implemented an artificial agent with a hierarchical architecture that allows the acquisition of several skills and we tested its performance in a setup with continuous states and actions, comparing both PE-based and PEI-based IM signals generated by different mechanisms. Some of the tested systems are taken from the computational literature related to IMs, including both the works of other researchers and our own; others derive from existing mechanisms but have not been tested before. The origin of each mechanism is indicated in section 2.3 where the different algorithms are explained in detail.

2. MATERIALS AND METHODS

2.1. THE EXPERIMENTAL SETUP AND THE SIMULATED ROBOT

The experimental task (**Figure 1**) consists in learning to reach for different circular objects positioned within the work space of a simulated kinematic robotic arm. The system has to learn in the best way and possibly shortest time a certain number of different skills, solely on the basis of IM reinforcement signals.

There are 8 different objects, corresponding to 8 different tasks: 2 are easy to be learnt, 2 are difficult and 4 are impossible to reach. The difficulty of the tasks is estimated on the basis of preliminary experiments where we tested the average time needed by a non-modular system to learn each of the different tasks with a performance of 95% (which is the average target performance in our experiments): easy tasks only need less than 2000 trials to be learnt while difficult tasks need more than 20,000 trials. Note that what we needed was not the precise measure of the difficulty of each task, but two classes of tasks differing substantially in the amount of trials needed to be learnt.

The choice of presenting tasks with different degrees of complexity derives from the evidence that an agent (be it an animal, a human, or a robot) can try to learn a great number of different

abilities that typically vary considerably with respect to their learning difficulty, including many (probably the majority) that are not learnable at all (consider, for example, an infant trying to learn to reach for the ceiling). For this reasons, it is very important for a system to avoid trying to acquire unlearnable skills and to focus on those that can be learnt for the necessary amount of time (enough for a satisfying learning but no more than required).

The system is implemented as a simulated kinematic robot composed by a two degree-of-freedom arm with a “hand” that can reach for objects. The sensory system of the robot encodes the proprioception of the arm, i.e., the angles of the two joints. The output of the controller determines the displacement of the two joints in the next time step.

2.2. ARM CONTROLLER AND CODING

Since we are looking for a system able to learn different skills and cache them in its own repertoire of actions, we need an architecture where different abilities are stored in different components of the system (Baldassarre and Mirolli, 2013c). For this reason, the controller of the arm consists in a modular architecture (**Figure 2**) composed by n experts (8 in this implementation, one for each possible task) and a selector that determines which expert/task will be trained. For simplicity, we coupled each expert to a specific task so that the expert is reinforced only for reaching the associated object, but this assumption does not affect the generality of the results presented here.

Note that the values of the parameters in these experiments were chosen in different ways. The parameters of the experts are not directly connected to the goals of this work: here we are interested in which is the best IM signal for driving the acquisition of multiple skills regardless of the specific ability of the experts. For this reason, the parameters related to the experts are simply taken from our previous works (Santucci et al., 2010, 2013a; Mirolli et al., 2013). The parameters related to the selector and the selection procedure, as well as those connected to the reinforcement signal provided to the selector, derive from a hand search where we identified the values that guaranteed the best results. In particular, we isolated the crucial parameters (the

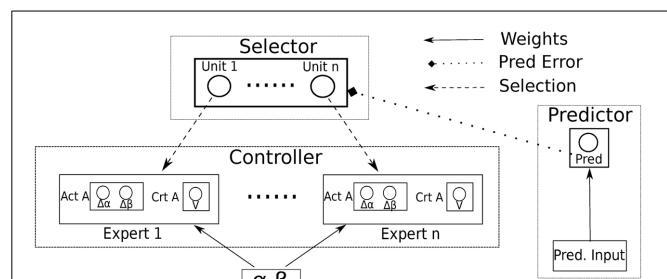
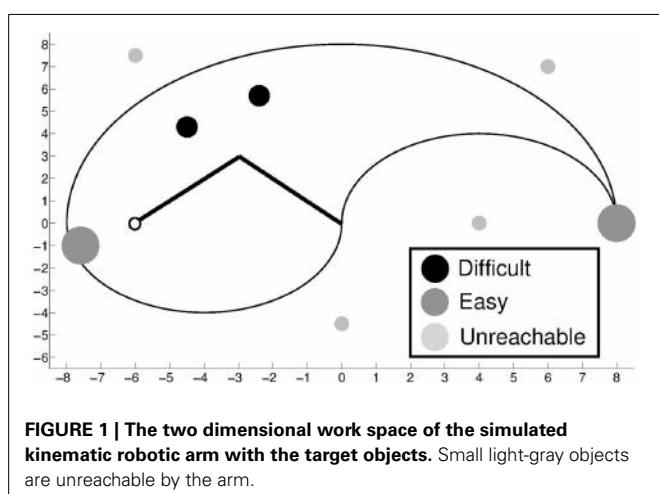


FIGURE 2 | The modular architecture of the system with the controller based on actor-critic experts, the selector and the predictor that generates the IM reinforcement signal driving the selector. n is the number of the tasks; Act A is the output of the actor of the expert, controlling the displacement of the joints of the arm in the next step; Crt A is the evaluation made by the critic of the expert.

learning rate of the predictors, the *temperature* of the *softmax* selection rule, and the temporal parameter α in the Q-learning rule that determines the activity of the unit of the selector: see below) and systematically (within limited ranges) changed their values in order to find a valid setup. Those that guarantee the best performance are the ones presented in the paper. Note that different values determine worse performances from a quantitative point of view (all the systems need more time to accomplish the tasks), but the differences between the experimental conditions are qualitatively stable.

Each expert is a neural network implementation of the actor-critic architecture (Barto et al., 1983) adapted to work with continuous state and action spaces (Doya, 2000). The input to the experts are the actual angles of the two joints of the arm, α and β (ranging in [0, 180]), coded through Gaussian radial basis functions (RBF) (Pouget and Snyder, 2000) in a two dimensional grid (10×10 units).

The evaluation of the critic (V) of each expert is a linear combination of the weighted sum of its input units. The actor of each expert has two output units, fully connected with the input, having a logistic transfer function:

$$o_j = \Phi\left(b_j + \sum_i^N w_{ji}a_i\right) \quad \Phi(x) = \frac{1}{1 + e^{-x}} \quad (1)$$

where b_j is the bias of output unit j , N is the number of input units, a_i is the activation of unit i and w_{ji} is the weight of the connection linking the input unit i to the output unit j . Each motor command o_j^m is determined by adding noise to the activation of the relative output unit:

$$o_j^m = o_j + q \quad (2)$$

where q is a random value uniformly drawn in $[-0.1; 0.1]$. The resulting commands are limited in $[0, 1]$ and then remapped in $[-25, 25]$ and control the displacement of the related arm joint angles.

In each trial, the expert that controls the arm is trained through a TD reinforcement learning algorithm. The TD-error δ is computed as:

$$\delta = (R_e^t + \gamma_k V^t) - V^{t-1} \quad (3)$$

where R_e^t is the reinforcement for the expert e at time step t , V^t is the evaluation of the critic of the expert at time step t , and γ is a discount factor set to 0.9. The reinforcement is 1 when the hand touches the object associated with the selected expert, 0 otherwise.

The connection weight w_i of critic input unit i is updated in the standard way (Sutton and Barto, 1998):

$$\Delta w_i = \eta^c \delta a_i \quad (4)$$

where η^c is a learning rate, set to 0.08.

The weights of each actor are updated as follows (see Schembri et al., 2007a):

$$\Delta w_{ji} = \eta^a \delta (o_j^m - o_j) (o_j (1 - o_j)) a_i \quad (5)$$

where η^a is the learning rate, set to 0.8, $o_j^m - o_j$ is the discrepancy between the action executed by the system (determined by adding noise) and that produced by the controller, and $o_j (1 - o_j)$ is the derivative of the logistic function.

The selector of the experts is composed by n units, one for each expert/task to be selected/learnt. At the beginning of every trial the selector determines the expert controlling the arm during that trial through a *softmax* selection rule (Sutton and Barto, 1998). The probability of unit k to be selected (P_k) is thus:

$$P_k = \frac{\exp \frac{Q_k}{\tau}}{\sum_{i=0}^n \exp \frac{Q_i}{\tau}} \quad (6)$$

where Q_k is the *Q*-value of unit k and τ is the *temperature* value that rescales the input values (here the *Q*-values) and so regulate the noise of the selection.

The activity of each unit is determined by a *Q-learning* rule used to cope with n-armed bandit problems with non-stationary rewards Sutton and Barto (1998):

$$Q_k^{t+1} = Q_k^t + \alpha [R_s^t - Q_k^t] \quad (7)$$

where Q_k^t is the *Q*-value of the unit corresponding to the selected expert during trial t , α is a temporal parameter set to 0.35 and R_s^t is the reinforcement signal obtained by the selector.

The reinforcement signal (R_s^t) driving the selection of the experts is the intrinsic reinforcement that we want to analyse in order to find the one that is the most suitable for autonomously learning multiple skills. Such signal is based on the error, or the improvement in the error, of a predictor of future states of the world. We now consider the different signals compared in this work.

2.3. IM SIGNALS AND PREDICTORS

2.3.1. Prediction error signals

As mentioned in section 1, we tested the IM signals and the mechanisms (predictors) implemented to generate such signals that are most used in the literature on IMs (see Figure 3 for a scheme of the different experimental conditions).

IM Type	Signal Type	Prediction Error (PE)		Prediction Error Improvement (PEI)	
		Training	Input	Standard	TD
Knowledge Based (KB)	State-Action-Predictor (SA)	KB-PE		KB-PEI	
		SAP-PE	SAP-TD-PE	SAP-PEI	SAP-TD-PEI
	State-Predictor (S)	SP-PE	SP-TD-PE	SP-PEI	SP-TD-PEI
	Only Task-Predictor (T)	TP-PE		TP-PEI	

FIGURE 3 | Scheme of the different experimental conditions, divided by typology of signal, typology of intrinsic motivations, input, and training algorithm. Note that the random (RND) condition is not mentioned in this table because it does not use any reinforcement signal to determine the selection of the experts. See Section 2.3.1 and 2.3.2 for a detailed description of all the different conditions.

- **Knowledge-Based Predictor (KB-PE):** The first IM reinforcement signal was the prediction error (PE) of a predictor of future states of the world (Schmidhuber, 1991a): in this model, the IM signal is represented by the absolute value of the error in predicting future states. The proposed mechanism was based on a forward model receiving the actual state and the planned action as input and predicting the next state. The idea is that the system, driven by the intrinsic PE signal, would explore the environment looking for new states that are not predictable by the forward model, acquiring at the same time the competence in new skills related to those states.

However, such predictors generate a signal which is coupled to the knowledge of the mechanism (learning the model of the world) and not to the competence of the system (learning skills). This signal can be considered as a purely knowledge-based prediction error (KB-PE) IM signal which may turn out to be inadequate for driving the acquisition of a repertoire of skills (see Santucci et al., 2012a; Mirolli and Baldassarre, 2013). In order to provide a stronger link between the predictor and the competence of the system, an effective solution is to change the target of the predictions. Instead of trying to anticipate every possible future configuration, the predictor has to anticipate only one particular state, the one connected to the trained skill, i.e., the *goal state*. In this way the PE signal is generated on the basis of the error in predicting the achievement of the goal, i.e., the generation of the final result of the skill that the agent is learning. Unlike KB-IM, this kind of signals can be considered competence-based (CB) IM signals and the predictors that generate them can be identified as CB-IM mechanisms (for the distinction between KB-IM and CB-IM, see also Oudeyer and Kaplan, 2007b).

Here we tested different CB-IM mechanisms. While all these mechanisms learn to predict the achievement of the goal state, they differ in the information received as input. Note that all the predictors also receive the information on which expert/task is currently trained by the system.

- **State-Action Predictor (SAP-PE):** This predictor has the same input as KB-PE mechanism, that is the actual state (the two joints of the arm, α and β) and the planned action ($\Delta\alpha$ and $\Delta\beta$), coded through RBFs. Training follows a standard delta rule. Examples of SAP-PE can be found in Santucci et al. (2010, 2013b).
- **State Predictor (SP-PE):** The SP-PE is not widespread in the literature. A similar predictor can be found in Barto et al. (2004), although this work proposed a system implemented within the option theory framework (Sutton et al., 1999), where the focus is more on the learning of the deployment of previously acquired skills rather than on the learning of the skills themselves. In our previous works (Santucci et al., 2012a, 2013b) we found that because its input is composed only by the actual state of the agent this kind of predictors are more closely coupled to the competence of the system than the SAP-PE: SP-PE mechanism is able to anticipate the achievement of the goal only when the agent has learnt the correct actions from the different states. Input is coded through RBFs. SP-PE is trained through a standard delta rule.

- **Temporal Difference SAP (SAP-TD-PE):** This predictor has the same input as SAP-PE but it is trained through a TD-learning algorithm with a discount factor set to 0.99. The implementation of this mechanism derives from the knowledge acquired in previous works (Mirolli et al., 2013; Santucci et al., 2013a) where we found that standard SAP-PE predictors do not work well with continuous states and actions. Providing the predictors with a TD algorithm solves some of these problems (for a generalization of TD-learning to general predictions, see Sutton and Tanner, 2005).
- **SP-TD-PE:** As for the SAP-TD-PE mechanisms, this predictor is the TD-learning version of SP-PE.
- **Task Predictor (TP-PE):** This predictor is inspired by our work in a simple grid-world scenario (Santucci et al., 2012a). A similar mechanism is implemented also in Hart and Grupen (2013). Differently from all the previous predictors, TP-PE does not make step-by-step predictions but a single prediction, at the beginning of the trial, on the achievement of the selected task. The input of this predictor consists only of the task/expert that has been selected, encoded in a n -long binary vector, with n equal to the number of tasks. The predictor is trained through a standard delta rule. These characteristics should provide a complete coupling between the signal generated by the predictor and the competence of the system in achieving each task: the predictor has no further information and can learn to anticipate the achievement of the target state only when the agent has really acquired a high competence in the related skill. In this way the selector should give the control to an expert only when it is effectively learning, shifting to a different expert when the competence to perform the related task has been completely acquired.

All CB-PE mechanisms generate a prediction (P) in the range [0, 1] related to the expectation that the system will accomplish the goal state within the time out of the trial. The error in predicting the goal state provides the intrinsic reinforcement signal to the selector of the system, whose activity determines which expert controls the system during the next trial and, at the same time, determines the expert that is trained by the system. This PE reinforcement signal is always positive: with the KB mechanism it is equal to the absolute value of the error; with CB mechanisms it is $1-P$ when the system reaches the goal state and 0 otherwise.

For all the systems implemented with the different PE mechanisms, the *temperature* τ value of Equation 6 is set to 0.01.

2.3.2. Prediction error improvement signals

As pointed out by Schmidhuber (1991b), PE signals may encounter problems in stochastic environments: if the achievement of a target state is probabilistic, the predictor will continue to make errors indefinitely. This means that the reinforcement will be never completely canceled and the system may keep on trying to train a skill even when it cannot improve any more. In order to solve this problem several systems (e.g., Schmidhuber, 1991b; Oudeyer et al., 2007a) use the improvement of the prediction error (PEI) rather than the PE as the IM signal.

For this reason, we also tested all the mechanisms described in section 2.3.1 using their PEI (instead of the PE) as the

reinforcement signal for the selector. Examples of KB-PEI can be found in Schmidhuber (1991b); Huang and Weng (2002); Baranes and Oudeyer (2009); an example of a SAP-PEI mechanism can be found in Oudeyer et al. (2007a). All the other mechanisms (SP-PEI, SAP-TD-PEI, SP-TD-PEI and TP-PEI), are tested here for the first time.

The PEI at time t was calculated as the difference between the average absolute PEs calculated over a period T of 40 time steps:

$$PEI_t = \frac{\sum_{i=t-(2T-1)}^{t-T} |PE|_i}{T} - \frac{\sum_{i=t-(T-1)}^t |PE|_i}{T} \quad (8)$$

In addition to the other mechanisms, in the PEI condition we also tested another CB-IM signal (Schembri et al., 2007a,b; Baldassarre and Mirolli, 2013b):

- **Temporal-Difference Predictor (TD):** This mechanism uses the TD-error (see Equation 3) of the selected expert as the intrinsic reinforcement signal that drives the selector. More precisely, here we use the average TD-error within the trial as the IM signal. Indeed, the TD-error can be considered a measure of the expert improvement in achieving its reinforcement and for this reason a measure of the competence improvement.

For all the systems implemented with the different PEI mechanisms the *temperature* τ value of the Equation 6 is set to 0.008. For the TD mechanism, the *temperature* τ is 0.01, while the α of Equation 7 is 0.25.

In order to better evaluate the performance of the simulated robot in the experimental setup when driven by the IM signals generated by the different mechanisms, we also tested a system that selects experts randomly (RND). Sometimes random strategies can indeed turn out to be surprisingly good: however, the best IM signal to drive the selection and acquisition of different skills in the shortest time, should guide the system better than a random selection.

2.4. HYPOTHESES AND COMPARATIVE CRITERIA

The main purpose of this work is to investigate which is the most suitable IM learning signal for driving the acquisition of a repertoire of different skills in the shortest time. In our previous works (Santucci et al., 2012a, 2013b), we proposed that the most important feature of such a signal should be its coupling with the competence in the skill that the system is trying to learn. For this reason our first hypothesis is that competence based signals should perform better than knowledge based ones.

With respect to the various CB mechanisms implemented, we expect that the TD versions of SAP and SP conditions should perform better than their normal versions since we know from previous works (Mirolli et al., 2013; Santucci et al., 2013a) that the latter ones do not work well with continuous states and actions. Furthermore, we also expect TP to perform better than both SAP and SP. With respect to PE vs. PEI, we predict that PE signals may behave a bit better than PEI signals, as the latter are probably more noisy and less strong than the former. Finally, we do not know how the TD error signal may perform with respect to the other PEI signals.

We compare the different IM signals by measuring their velocity in learning multiple tasks. In particular, we run different experiments (see section 3) and count the number of trials (averaged over several repetitions of the experiment) needed by each condition to achieve an average performance of 95% in the 4 learnable tasks. We chose the average of 95% as the target performance since we want a value that is able to identify a satisfying capability of a system to learn different skills. If we used a different target performance (e.g., 90 or 99%) they would be qualitatively the same.

3. RESULTS

Each condition was tested for 400,000 trials. At the beginning of every trial the selector determines which expert will control the activity of the arm in that trial. Each trial ends if the selected expert reaches its target object or after a time out of 20 time steps.

For every mechanism, we ran different simulations varying the learning rate (LR) of the predictor (9 different values) because we wanted to be sure that the results were not dependent on the use of a specific set of LRs. For each LR we ran 20 repetitions of the experiment. In the TD and RND condition, where there is not a separate predictor (in RND there is no IM signal, in TD we use the TD-error of the experts), we ran 180 repetitions of the experiment to balance the total number of replications in the other conditions.

3.1. PE SIGNALS

Figure 4 shows the number of trials (averaged over the 180 replications) needed by the different PE conditions to achieve an average performance of 95% in the 4 learnable tasks. The results clearly underline, confirming one of our hypotheses, how the TP-PE mechanisms is the one that generates the best signal to drive the system in achieving a high average performance in the learnable tasks in the shortest time (average of about 130,000 trials). As expected (see Section 2.3.1 and 2.4), the SAP-PE and the SP-PE are not able, working within continuous states and actions,

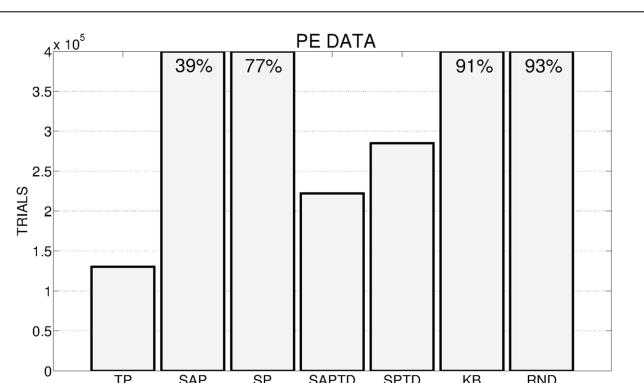


FIGURE 4 | Average number of trials needed by the different conditions to achieve an average performance of 95% in the 4 learnable tasks (average results of 180 replications: 20 replication by 9 learning rates for the systems with predictors, 180 replications for the random system) in the different experimental conditions. If a system has not reached 95% at the end of the 400,000 trials we report on the corresponding bar the average performance at the end of the simulation.

to generate a good signal to guide the selection and the learning of skills. SAP-TD-PE and SP-TD-PE are able to drive the system in achieving the average target performance within the 400,000 trials but they are slower than the TP-PE system. Both KB-PE and RND conditions can reach high performance within the end of the experiment (more than 90%), but they are not able to achieve the target value of 95%. An interesting result is that the system driven by the random selection reaches an average performance (93%) higher than the one driven by KB-PE mechanism (91%).

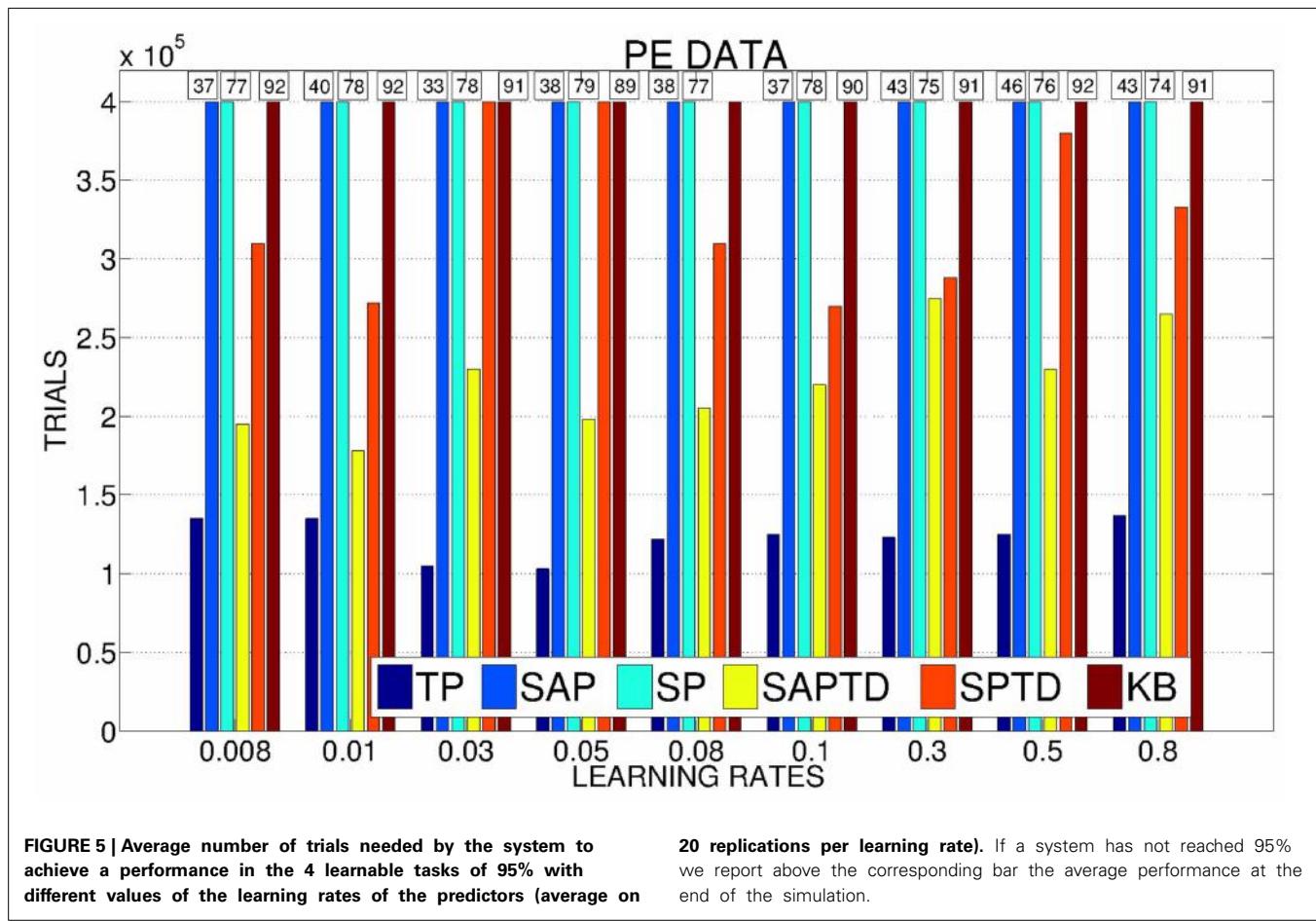
In **Figure 5** we show a detailed analysis of the average performance of the system in the different conditions with different values of the learning rate for the predictors. SAP-PE and SP-PE are not able, regardless from the learning rate of the predictor, to achieve the target performance, while SAP-TD-PE and SP-TD-PE seems to be sensitive to the value of the learning rate of the predictor (SP-TD-PE more than SAP-TD-PE). Differently, TP-PE is very robust with respect to the value of the learning rate of the predictor: regardless of this value this condition is always the best performer, being able to achieve a high performance in a short time.

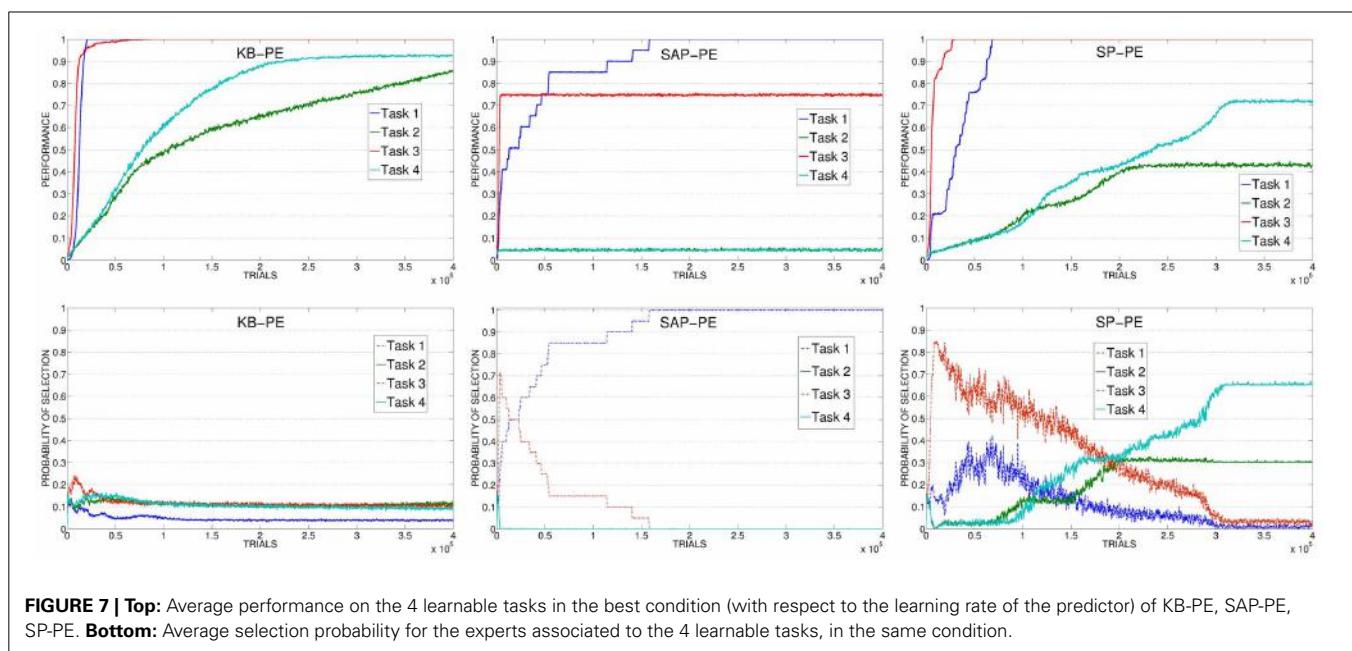
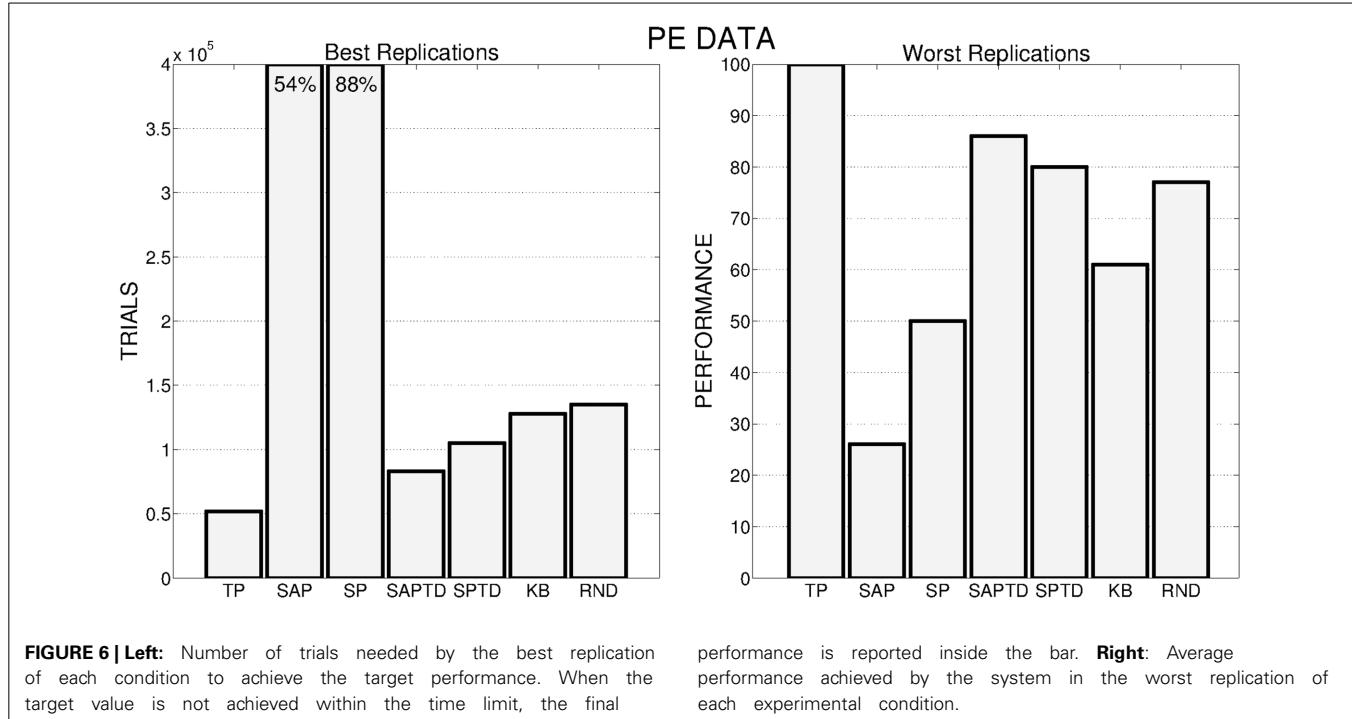
These general results are even more evident if we look at **Figure 6**, where the performance of the best and worst replications of every condition are shown: the overall best performance is achieved by a replication of the TP-PE condition that is able to reach the target performance in about 50,000 trials. As in the case of average

performances, the best replications of SAP-PE and SP-PE are not able to reach the target performance while KB-PE and RND have comparable performance. Even more impressive are the results of the worst replications: the TP-PE mechanism is the only one that is able to drive the system in achieving the target performance within the given time also in its worst replication. The other conditions reflect the average results, with the KB-PE condition performing worse than random selection in its worst replication.

To understand the causes of these results, for each condition we analyzed the average selections of the experts connected to the 4 learnable tasks during time and the average level of performance achieved on those tasks. Data are related to the best learning rate value of the predictor of each different condition. In this way we can check if the signal generated by the predictors is able to drive the selector in a proper way, following the actual competence acquired by the experts. Data of RND system are not shown: in this case experts are always selected (on average) uniformly, and hence the system wastes time in selecting experts that cannot learn anything or that have already learnt their tasks (e.g., the two easy tasks).

Figure 7 (left) shows the results of the KB-PE mechanisms. In this condition the system is not driven by an IM signal connected to the competence of the system in learning the skills, but to the knowledge acquired by the predictor in anticipating every possible future state. For this reason the system is not selecting the





experts connected to the tasks that are still to be learnt, but rather the experts that are surprising the predictor reaching whatever unpredicted state. These experts include also those related to the 4 non-learnable tasks. This process leads to a random selection (random-selection value is 0.125 because it is calculated on all the 8 tasks). While this is not a problem for the two easiest tasks (task 1 and task 3) that are learnt after few trials, the canceling of the IM signal and the consequent absence of a focused learning severely impairs the learning of the difficult tasks (task 2 and task 4).

The result of the KB-PE condition confirm one of our main hypotheses, clearly underlining how a KB-IM signal is inadequate to properly drive an agent in learning different skills: it either continues to select already learnt tasks, or it does not properly select those that are still to be learnt. This is the reason why, if we are looking at improving the competence of a system, we should use CB-IM mechanisms.

If we look at data related to SAP-PE and SP-PE (**Figure 7**, center and right) it is clear that these mechanisms are not able to

cancel in a proper way the PE signal provided by the achievement of the goal states. For this reason SAP-PE, on average, focuses on one of the easiest tasks (whose target states, on average, are rapidly discovered by the system) although the robot has completely acquired the related competence. SP-PE is able to anticipate the achievement of the easy tasks, but it learns too slowly these predictions: for this reason, although task 1 and task 2 have both been learned at about 70,000 trials the system still focuses on them for further trials, wasting precious time for learning the more difficult skills.

SAP-TD-PE and SP-TD-PE (Figure 8, left and center) present the opposite problem: these mechanisms learn very fast to predict the reaching of the objects, even faster than the actual competence of the system in those tasks. Although these are CB mechanisms, the learning process of these predictors is not strictly coupled with the ability of the system to reach for the objects. This is evident comparing the progress in the performance with the selections: the predictors cancel the signals before the system has acquired the competence related to the different tasks determining a selection which is not optimally coupled to the actual performance of the system. However, in spite of this problem, these mechanisms are able to guide the system in reaching the target performance within a reasonable time. This is because, differently from KB-PE and RND, although turning too fast to a random selection, they perform selections only on the 4 learnable tasks (that are the only ones that can generate a PE) and not on all the 8 tasks. SAP-TD-PE and SP-TD-PE do not provide a perfect IM signal, but they are a good example of how even a sub-optimal CB-IM signal is able to drive the learning of skills better than a KB signal.

Differently from all the other conditions, the TP-PE mechanism (Figure 8, right) is able to drive the complete learning of the skills in relatively few trials. The reason of this performance is connected to the signal generated by the TP-PE mechanism: this signal is strictly coupled with the competence of the system in the

task that it is learning. Looking at the average development of the experiment, it is clear how the selector, driven by this CB-IM signal, assigns the control of the robot only to an expert connected to a task that has still to be learnt, shifting to another one when a skill has been fully achieved. Easy skills need just few trials to be learnt and for this reason the system focuses on their training (and selection) only for a very short time at the beginning of the experiment. As soon as the predictor has learnt to anticipate the achievement of those target states, it cancels their respective signals and drives the agent to search for other skills to acquire. Difficult tasks require a longer time to be learnt so the system focuses on selecting the related experts longer, until a high performance has been achieved. When all the tasks have been learnt the predictor has learnt to anticipate the achievement of all the target states, so the selector receives no more intrinsic reinforcements and generates an (almost) random selection.

3.2. PEI SIGNALS

Figure 9 shows the average number of trials needed by the system to achieve the target performance of 95% within the different conditions. As with the PE signal, also with the PEI signal the TP-PEI condition is the one that is able to guide the system in achieving the target performance in the shortest time. However, the average number of trials needed by those conditions that best perform with PE signals (TP, SAP-TD, SP-TD) is raised. At the same time, those conditions that with PE signal were not able to achieve the target average performance (95%) in the learnable tasks, with PEI significantly improve their results, with SAP-PEI and SP-PEI reaching a performance similar to SAP-TD-PEI and SP-TD-PEI. This is due to the properties of PEI signal: if a predictor is not able to improve its ability to anticipate the achievement of a target state, there is no improvement in the prediction error and the signal is canceled. So, despite the predictor is not able to correctly anticipate the achievement of the easy tasks even when

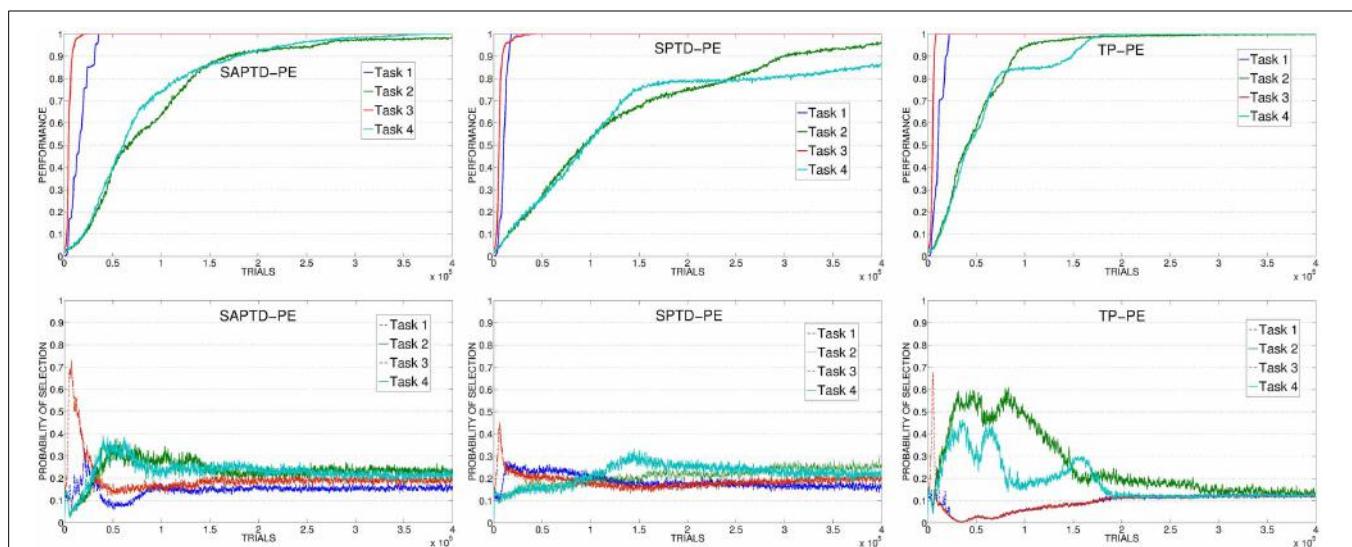


FIGURE 8 | Top: Average performance on the 4 learnable tasks in the best condition (with respect to the learning rate of the predictor) of SAP-TD-PE, SP-TD-PE, TP-PE. **Bottom:** Average selection probability for the experts associated to the 4 learnable tasks, in the same conditions.

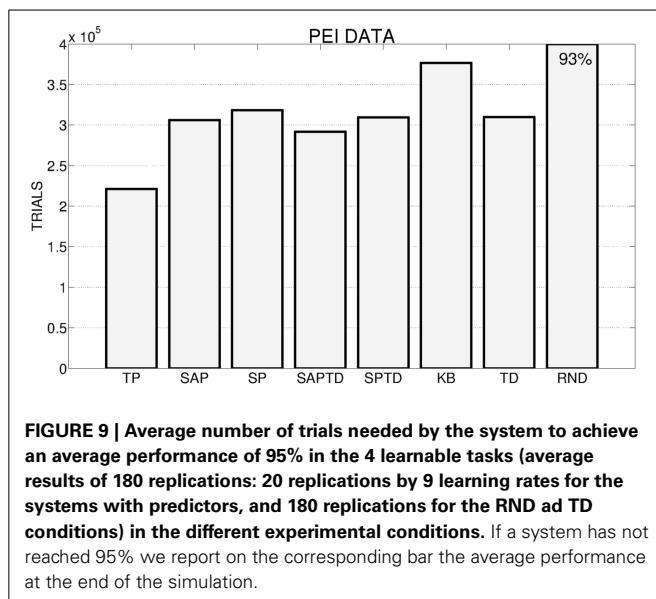


FIGURE 9 | Average number of trials needed by the system to achieve an average performance of 95% in the 4 learnable tasks (average results of 180 replications: 20 replications by 9 learning rates for the systems with predictors, and 180 replications for the RND ad TD conditions) in the different experimental conditions. If a system has not reached 95% we report on the corresponding bar the average performance at the end of the simulation.

their competence is fully acquired (as in SAP-PEI and SP-PEI conditions), the constant error generates no PEI signal and allows the system to shift to the selection of different experts possibly discovering new learnable skills. The TD condition guarantees a performance that is similar to those of the other CB signal (except for TP-PEI, which is the best performer), while when the system is driven by the KB-IM signal it is not able to achieve satisfying results: KB-PEI turns out to be the worst PEI condition.

As anticipated in our hypotheses, PEI signals are much noisier and weaker than PE signals. This is clear from **Figure 10**, showing how all the conditions (including TP) present a high sensitivity to the variation in the learning rate of the predictors. However, TP-PEI is the one that is able to drive the system in achieving the target performance in the shortest number of trials (only 150,000, on average, with learning rate 0.05).

Data on the average performances are confirmed by **Figure 11**, where we show the best (**Figure 11**, left) and worst (**Figure 11**, right) replications of all the different conditions. As for PE signal, also with PEI the best replication of the TP-PEI condition is the absolute best among all the replications of all the conditions and even its worst replication is the one that reaches the highest performance compared to the worst replications of the other conditions. KB-PEI confirms to be the worst PEI condition: even its best replication (**Figure 11**, left) is performing as the RND selector. TD condition shows a great variance in its different replications: its best replication (**Figure 11**, left) is only the 5th performer, while its worst replication is the second best (among the worst replications of all the conditions) after the TP.

As with PE experiments (section 3.1), to better understand the results we analyzed data showing the average selections of the experts connected to the 4 learnable tasks during time and the average level of performance achieved on those tasks. Data are related to the best learning rate value of each different condition, while for TD condition we look at the average performance and

selections on 20 replications (consecutive and including the best replication of the condition).

The poor performance of KB-PEI (**Figure 12**, left) is related to the bad selection determined by the KB-IM signal: the experts related to the 4 learnable tasks are clearly selected randomly.

When driven by CB-IM signals the system reaches a better performance, with differences between the conditions implemented with different mechanisms. In SAP-PEI and SP-PEI conditions the selection is very noisy (**Figure 12**, center and right). Although learnable tasks are selected more than in RND and KB conditions, the already weak signal is flattened by the activity of the predictors that are not able to significantly improve in their ability to anticipate the target states.

SAP-TD-PEI and SP-TD-PEI (**Figure 13**, left and center) are able to cancel the signal deriving from the rapidly learnt easier tasks, but at the same time they present the problem we found with the PE: these mechanisms can be too fast in canceling the IM signal, determining a decrease in the probability of selecting the complex tasks even if the system has still competence to acquire. This is confirmed by looking at data of SAP-TD-PEI condition, where the PEI signal for task 4 is drastically decreased around 200,000 trials, when the system has reach an average performance on that task of only about 80%.

As in the experiment with the PE signal, the TP-PEI mechanisms is the one that is able to drive the system in selecting and learning the different skills in the shortest time. The reason is the same as with PE results: even in its PEI version, the CB-IM signal generated by the TP mechanism is the only one that is closely connected to the competence acquired by the system in the different learnable tasks (**Figure 13**, right). Easy tasks, which are learnt very fast, are selected only during the short time needed to raise their performance. Thanks to the canceling of the intrinsic reinforcement signal provided to the selector, the system is able to shift to the complex tasks. At about 150,000 trials, on average, the system has reached a high performance on task 4: due to the connection of the TP mechanism to the competence of the agent, the PEI-IM signal related to that task fades away and the system focuses only on the skill that at that time of the experiment is the least efficient (task 2).

As mentioned in section 2.3.2, together with the different PEI signals we also tested another CB-IM signal provided by TD-error of the selected expert. As previously described, the average performance of TD condition is similar to those of other CB-IM conditions with PEI signal (except for TP, which is the best performer). However, if we look at the average performance on 20 replications (consecutive and including the best replication of this condition) we can see that when driven by the TD signal the system reaches a performance that is similar or even better than those of the other conditions (except for TP) in their best learning rate condition (confront **Figure 14**, left, with **Figures 12, 13**, top). Indeed, if we look at the average selections (**Figure 14**, right), we can see that TD signal is able to generate a sequence of selections that are connected to the competence progress of the system, although less than the one provided by the TP mechanism.

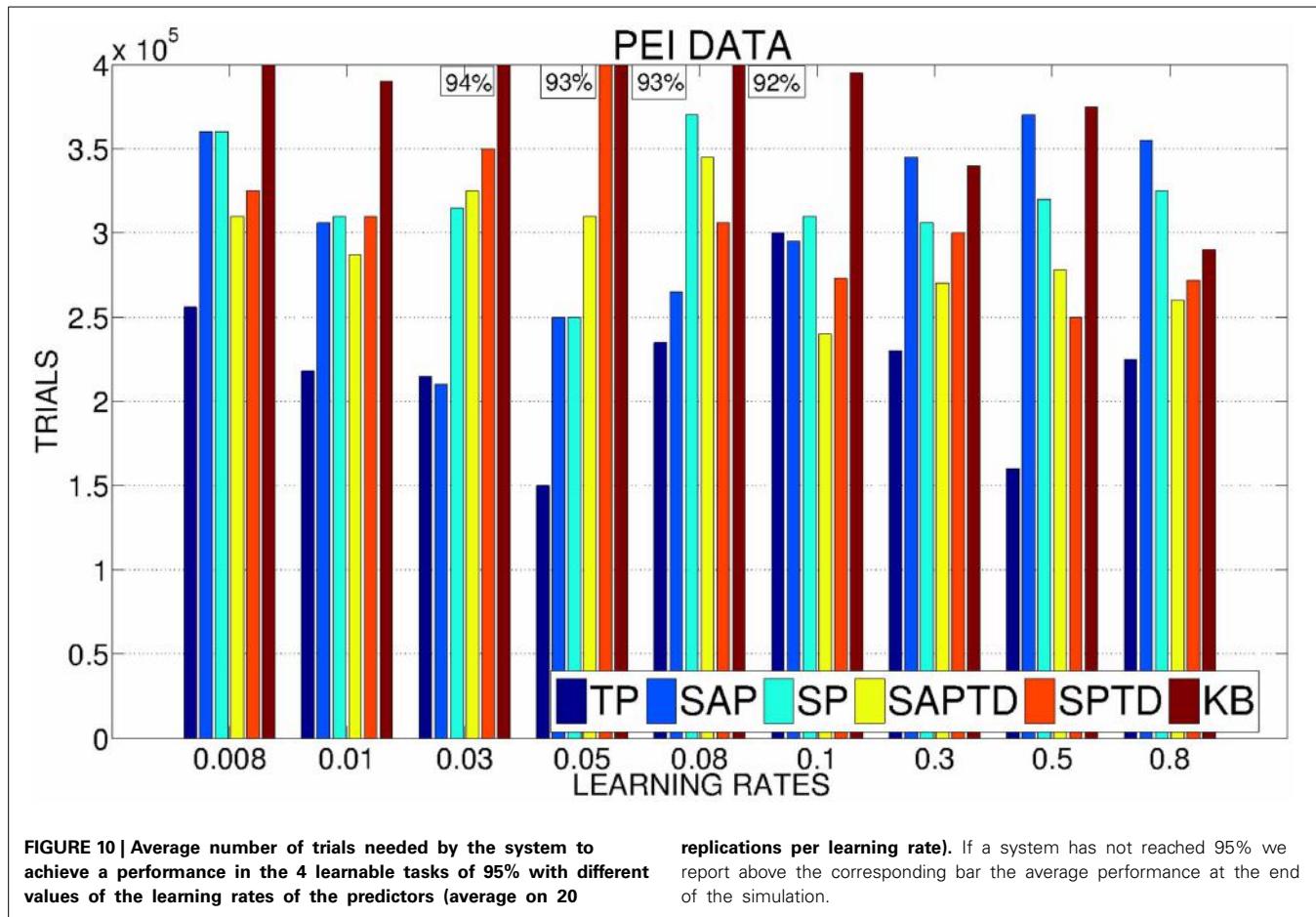


FIGURE 10 | Average number of trials needed by the system to achieve a performance in the 4 learnable tasks of 95% with different values of the learning rates of the predictors (average on 20

replications per learning rate). If a system has not reached 95% we report above the corresponding bar the average performance at the end of the simulation.

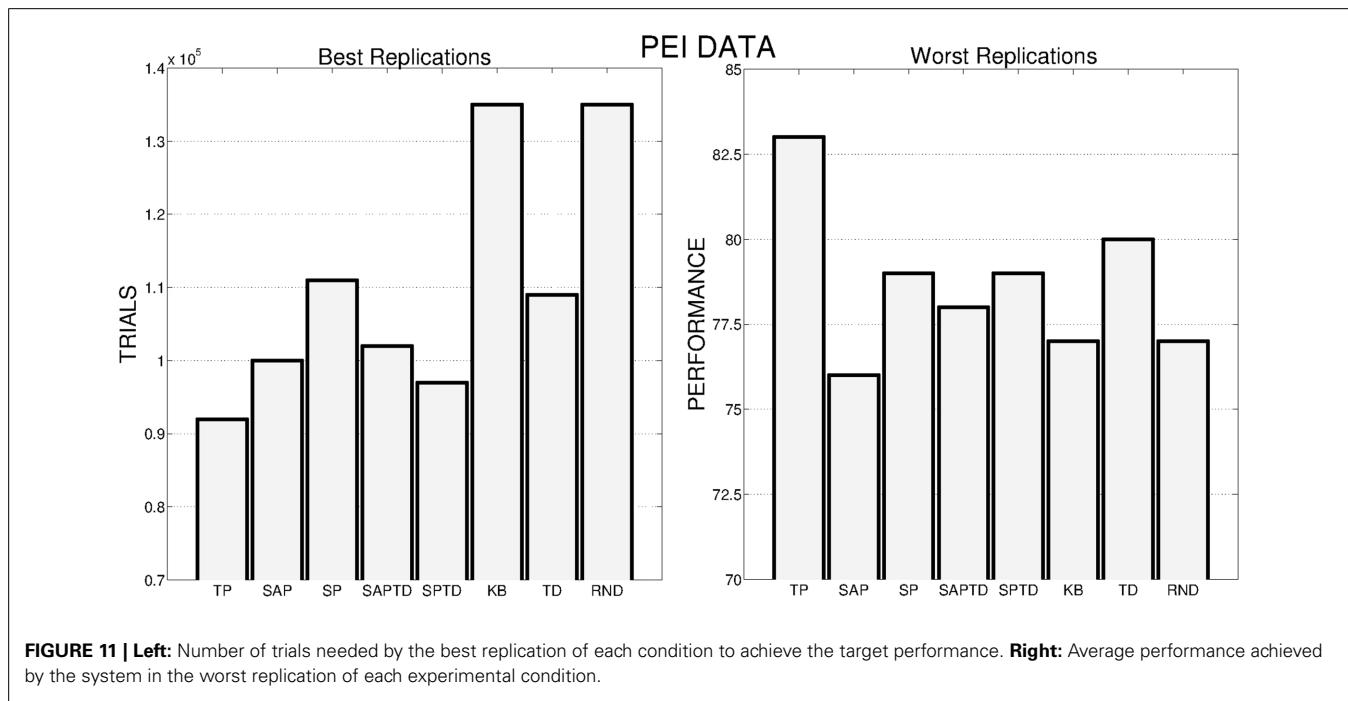


FIGURE 11 | Left: Number of trials needed by the best replication of each condition to achieve the target performance. **Right:** Average performance achieved by the system in the worst replication of each experimental condition.

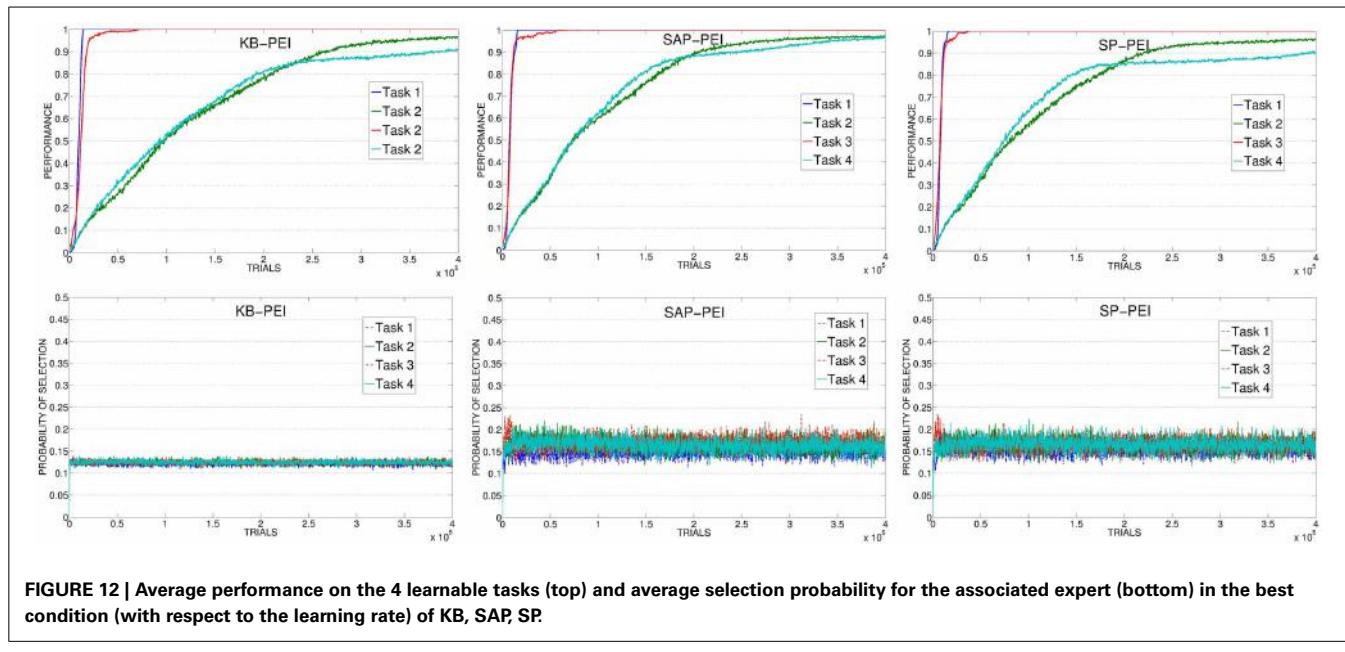


FIGURE 12 | Average performance on the 4 learnable tasks (top) and average selection probability for the associated expert (bottom) in the best condition (with respect to the learning rate) of KB, SAP, SP.

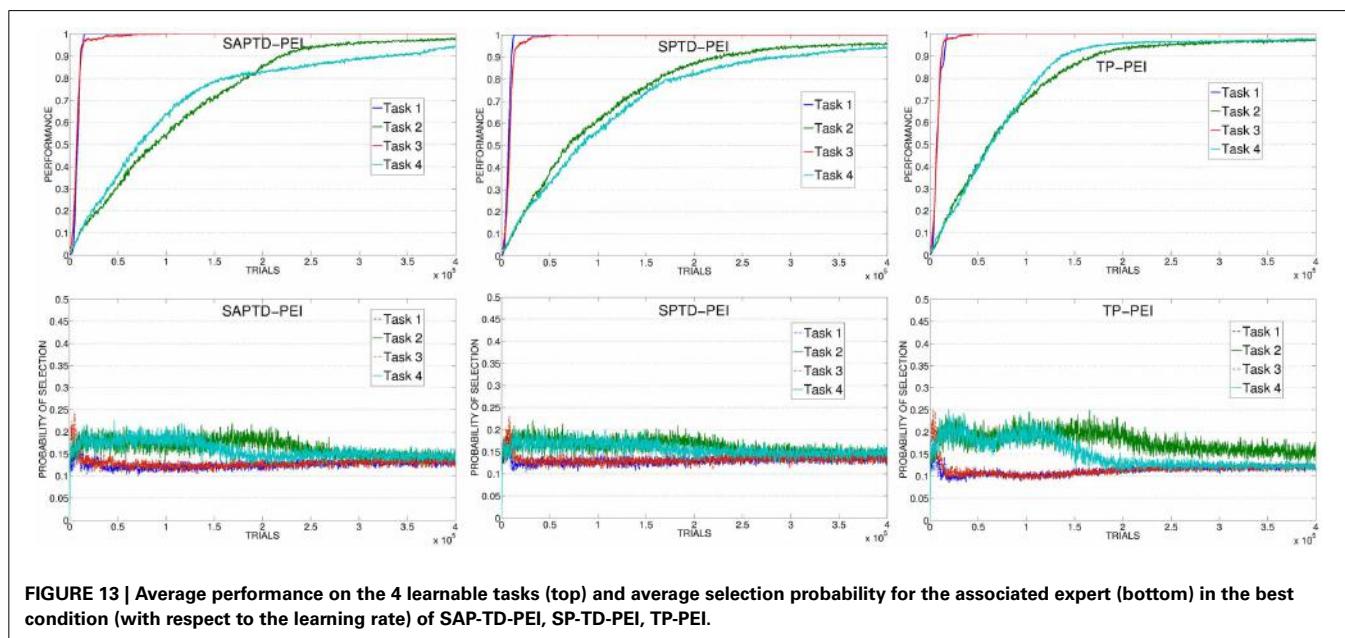


FIGURE 13 | Average performance on the 4 learnable tasks (top) and average selection probability for the associated expert (bottom) in the best condition (with respect to the learning rate) of SAP-TD-PEI, SP-TD-PEI, TP-PEI.

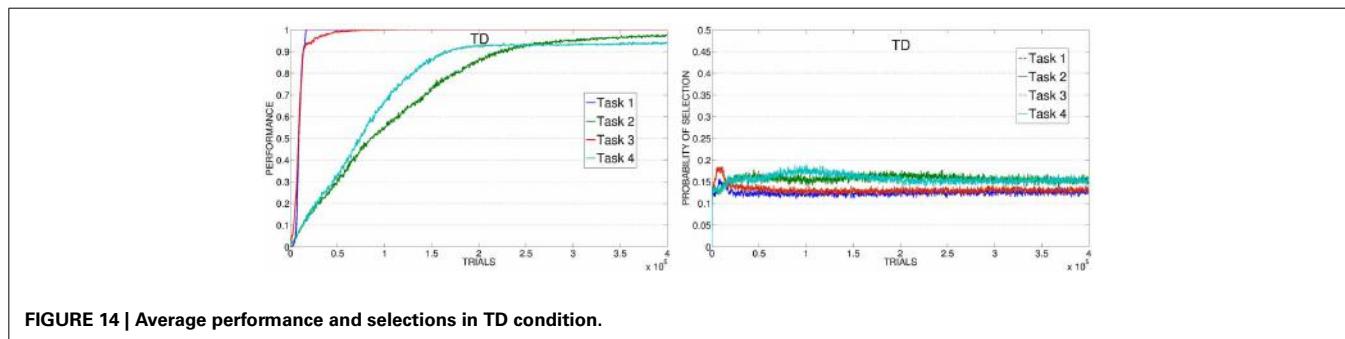


FIGURE 14 | Average performance and selections in TD condition.

4. DISCUSSION

In this paper we analyzed different kinds of IM signals in order to find the most suitable to drive a system in selecting and learning different skills in the shortest time. To tackle this important issue, we implemented a simulated two-dimensional kinematic robotic arm with a hierarchical architecture able to train and cache different skills and we tested it within continuous spaces and actions in an experimental scenario where the agent had to learn to reach different objects.

The first important result validate one of our main hypotheses: a purely KB-IM signal (as those implemented in Schmidhuber, 1991a,b; Huang and Weng, 2002) is not able to satisfactorily drive the acquisition of multiple skills. This signal is coupled to the knowledge of the KB predictor that tries to anticipate every possible future state of the world. The PE or PEI signal deriving from this kind of mechanism drives the system in exploring the environment without any specific target: this is why the performance of the KB condition is similar to RND condition, where the system is guided by a random selection of its experts. Note that the implementation provided in this work helps the KB mechanisms. Indeed, here we used the intrinsic reinforcement signal to drive the selection of the experts. In a previous work (Santucci et al., 2012a), we showed that if the KB-IM signal is provided directly to an actor-critic expert the system continues to explore the environment to train the predictor without learning any skill. With our results we are not saying that KB-IM are useless or wrong: simply they are involved in different processes, which are related to knowledge acquisition more than competence acquisition.

In order to optimize the IM-based acquisition of skills, learning signals have to be strictly connected to the actual competence in those skills, i.e., to the actual competence in achieving target goals. CB-IM signals provide such a coupling and the results of our experiments underlie how the stronger that coupling, the better the performance of the system (see **Figure 15** for the ranking of the results of all the experimental conditions). Indeed, not all the CB-IM mechanisms guarantee the same close connection between the correctness of the predictor and the competence acquired by the system. Some mechanisms like SAP and SP (especially when generating a PE signal) are not good predictors in continuous spaces and actions as they are *too slow*: they are not able to properly cancel the IM signal even if the agent has fully acquired the related competence, thus leading the system to focus on already trained experts. Other CB mechanisms (SAP-TD, SP-TD) turned out to provide a useful learning signal for the acquisition of skills, although they present the problem of being too fast in canceling the intrinsic reinforcement signal that fades away before the robot has completely learnt the related skills.

As expected, the condition that was able to learn all the skills in the shortest time, both in PE and PEI conditions, was the one where the IM reinforcement signal for the selector was generated by what we called TP mechanism: a predictor of the goal states (the target states connected to the different skills) that receives as input only the information on which expert has been selected to be trained. The mechanism that we proposed provides a close connection between the ability of the predictor in anticipating future target state and the actual competence acquired by the agent in the related skill. This coupling guarantees an IM signal

#Rank	PE	PEI	Average Performance	Performance Best Replication
1	TP-PE		130 k	52 k
2		TP-PEI	221 k	92 k
3	SAP-TD-PE		222 k	83 k
4	SP-TD-PE		285 k	105 k
5		SAP-TD-PEI	291 k	102 k
6		SAP-PEI	306 k	100 k
7		SP-TD-PEI	309 k	97 k
8		TD	310 k	109 k
9		SP-PEI	318 k	111 k
10		KB-PEI	376 k	135 k
11		RND	- (93%)	135 k
12	KB-PE		- (91%)	128 k
13	SP-PE		- (77%)	- (88%)
14	SAP-PE		- (39%)	- (54%)

FIGURE 15 | Ranking of the different experimental conditions summarizing the result of both PE and PEI signals with respect to the ability to reach the target average performance of 95% in the four learnable tasks. For every condition the performance of the best replication is also shown. Performances are measured in thousands of trials. If a condition has not reached 95% at the end of the 400,000 trials of the experiment we report the average performance at the end of the simulation.

which is particularly appropriate for the selection and acquisition of different skills: the intrinsic reinforcement is present when the system is learning a new task, it is canceled when the competence on that task has been learnt and reappears when a new, still-to-be-learnt task is encountered by the system.

Moreover, we also tested the TD condition where the TD-error signal of the active expert is used as the intrinsic reinforcement for the selector. This solution (Schembri et al., 2007a,b; Baldassarre and Mirolli, 2013b) is able to cope with the same problems connected to stochastic environments that may lead to use PEI signals instead of PE signals. The TD condition performs comparably to the other sub-optimal CB-IM driven conditions in PEI experiments. However, in its best replications, it is able to reach very high performance and, moreover, it presents important computational advantages: the absence of a separate component for the predictions reduces computational time and avoids the setting of its specific learning rate.

Despite the growing theoretical understanding of the differences between functions and mechanisms of IM (e.g., Oudeyer and Kaplan, 2007b; Stout and Barto, 2010; Santucci et al., 2012a; Mirolli and Baldassarre, 2013), their implications have not been fully exploited in specific models. In particular, there is still a confusion between KB mechanisms and CB mechanisms. Some still use KB-IM signals to drive the acquisition of competence, leading to inappropriate learning signals as underlined by the results of our present work. Others shifted, without realizing, to CB mechanisms probably because they encountered the problems connected to KB signals and competence acquisition. However, due to the lack of understanding of the differences between KB-IM and CB-IM, they turn out to implement sub-optimal CB mechanisms. An example is Oudeyer et al. (2007a) where, although they describe the implemented intrinsic signal as the

PEI of the knowledge of the system, they use the predictor to anticipate few (three) high-level abstract important states (visual detection of an object; activation of a biting sensor; perception of an oscillating object). These high-level states represent few relevant states among a huge number of non-interesting states, and each of them can be achieved only with sequences of actions. This predictor is very similar to the SAP we tested in our experiment, which in fact is a CB mechanism, even if its results are not the best possible.

Looking at the implementation of our system, a strong limit is the fact that the possible tasks to be learnt are given at the beginning of the experiment. A further step toward more autonomous and versatile agents would be to built systems that self-determine their goals. Recently, some effort has been made in the field of hierarchical reinforcement learning to find good solutions to the problem of setting useful goals. Most of these techniques (e.g., McGovern and Barto, 2001; Mehta et al., 2008; Konidaris and Barto, 2009) focus on searching adequate sub-goals on the basis of externally given tasks (reward functions). Only few works (e.g., Mugan and Kuipers, 2009; Vigorito and Barto, 2010) tried to implement systems able to set their own goals independently from any specific task, which is a fundamental condition for real open-ended autonomous development.

Another important point concerns the generality of our results. In future works it will be interesting to test the different IM learning signals in different experimental setups (e.g., adding more dimensions and degrees of freedom; using a dynamic arm) where different and possibly more difficult tasks have to be learnt: this would be a further confirmation of our results and conclusions. However, we believe that the main findings of this work are quite general. Indeed, the differences between KB-IM and CB-IM lie in the typology of information used to determine such signals and not on the specific setups they are implemented in. Similarly, the conclusion that a proper CB-IM mechanism has to generate a signal which is closely connected to the actual competence of the system is a general finding that can be exploited regardless of the particular architecture used to implement the agent.

Our expectation is that testing the different IM signals studied here in more realistic conditions will strengthen the advantages of using the TP signal with respect to the other implementations of IMs. In a real environment the number of skills that can be acquired is much larger than the one considered here, and the difficulty to learn the skills is much more heterogeneous. Moreover, in the real world there are strong dependencies between different competences, so that some skills can be learnt only after learning others. All these characteristics of real environments emphasize the importance for an IM signal to be strongly connected to the competence of the system, thus avoiding to waste time in easy (or previously learnt) tasks or in too difficult (or not possible) tasks, and focussing on the skills that can be learnt at the moment, which may be later exploited to learn other skills. Our results show that only a signal that is closely linked to the competence of the system is able to provide these general features.

Looking at the different typologies of IMs, our intuition is that they may play complementary roles, with KB-IM being able to inform the system of novel or unexpected states of the environment, driving the agent to generate new target states, and

CB-IM being able to guide the acquisition of the skills related to those targets. This further model, that tries to integrate the different typologies of IMs, will probably require a more complex architecture able to manage both the control of the effectors, the generation and selection of the different motivations and the combination of different IM learning signals.

ACKNOWLEDGMENTS

The authors want to thank Professor Andrew Barto and Bruno Castro da Silva for and fruitful discussions on the topics of this paper.

FUNDING

This research has received funds from the European Commission 7th Framework Programme (FP7/2007–2013), ÒChallenge 2 - Cognitive Systems, Interaction, Robotics, Grant Agreement No. ICT-IP-231722, project ÒIM-CLeVeR - Intrinsically Motivated Cumulative Learning Versatile Robots.

REFERENCES

- Baldassarre, G., and Mirolli, M. (eds.). (2013a). *Intrinsically Motivated Learning in Natural and Artificial Systems*. Berlin: Springer-Verlag. doi: 10.1007/978-3-642-32375-1
- Baldassarre, G., and Mirolli, M. (2013b). “Deciding which skill to learn when: temporal-difference competence-based intrinsic motivation (TD-CB-IM),” in *Intrinsically Motivated Learning in Natural and Artificial Systems*, eds G. Baldassarre and M. Mirolli (Berlin: Springer-Verlag), 257–278.
- Baldassarre, G., and Mirolli, M. (2013). *Computational and Robotic Models of the Hierarchical Organization of Behavior*. Berlin: Springer.
- Baranes, A., and Oudeyer, P. Y. (2009). R-iac: robust intrinsically motivated exploration and active learning. *IEEE Trans. Auton. Ment. Dev.* 1, 155–169. doi: 10.1109/TAMD.2009.2037513
- Baranes, A., and Oudeyer, P.-Y. (2013). Active learning of inverse models with intrinsically motivated goal exploration in robots. *Robot. Auton. Syst.* 61, 49–73. doi: 10.1016/j.robot.2012.05.008
- Barto, A., Singh, S., and Chantanez, N. (2004). “Intrinsically motivated learning of hierarchical collections of skills,” in *Proceedings of the Third International Conference on Developmental Learning (ICDL)* (San Diego, CA), 112–119.
- Barto, A., Sutton, R., and Anderson, C. (1983). Neuron-like adaptive elements that can solve difficult learning control problems. *IEEE Trans. Syst. Man Cybernet.* 13, 834–846. doi: 10.1109/TSMC.1983.6313077
- Berlyne, D. (1960). *Conflict, Arousal and Curiosity*. New York, NY: McGraw Hill. doi: 10.1037/11164-000
- Doya, K. (2000). Reinforcement learning in continuous time and space. *Neural Comput.* 12, 219–245. doi: 10.1162/089976600300015961
- Duzel, E., Bunzeck, N., Guitart-Masip, M., and Duzel, S. (2010). Novelty-related motivation of anticipation and exploration by dopamine (nomad): implications for healthy aging. *Neurosci. Biobehav. Rev.* 34, 660–669. doi: 10.1016/j.neubiorev.2009.08.006
- Harlow, H. F. (1950). Learning and satiation of response in intrinsically motivated complex puzzle performance by monkeys. *J. Comp. Physiol. Psychol.* 43, 289–294. doi: 10.1037/h0058114
- Hart, S., and Grupen, R. (2013). “Intrinsically motivated affordance discovery and modeling,” in *Intrinsically Motivated Learning in Natural and Artificial Systems*, eds G. Baldassarre and M. Mirolli (Berlin: Springer-Verlag), 279–300.
- Huang, X., and Weng, J. (2002). “Novelty and reinforcement learning in the value system of developmental robots,” in *Proceedings of the Second International Workshop Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems*, Vol. 94, eds C. Prince, Y. Demiris, Y. Marom, H. Kozima, and C. Balkenius (Lund: Lund University Cognitive Studies), 47–55.
- Kakade, S., and Dayan, P. (2002). Dopamine: generalization and bonuses. *Neural Netw.* 15, 549–559. doi: 10.1016/S0893-6080(02)00048-5

- Konidaris, G., and Barto, A. (2009). "Skill discovery in continuous reinforcement learning domains using skill chaining," in *Advances in Neural Information Processing Systems 22 (NIPS '09)*, eds Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta (Vancouver, BC), 1015–1023.
- Lopes, M., and Oudeyer, P. Y. (2012). "The strategic student approach for life-long exploration and learning," in *IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL)* (San Diego, CA). doi: 10.1109/DevLrn.2012.6400807
- McGovern, A., and Barto, A. G. (2001). "Automatic discovery of subgoals in reinforcement learning using diverse density," in *Proceedings of the Eighteenth International Conference on Machine Learning, ICML '01*, (San Francisco, CA: Morgan Kaufmann Publishers Inc.), 361–368,
- Mehta, N., Natarajan, S., Tadepalli, P., and Fern, A. (2008). Transfer in variable-reward hierarchical reinforcement learning. *Mach. Learn.* 73, 289–312. doi: 10.1007/s10994-008-5061-y
- Mirolli, M., and Baldassarre, G. (2013). "Functions and mechanisms of intrinsic motivations: the knowledge vs. competence distinction," in *Intrinsically Motivated Learning in Natural and Artificial Systems*, eds G. Baldassarre, and M. Mirolli (Berlin: Springer-Verlag), 49–72.
- Mirolli, M., Santucci, V. G., and Baldassarre, G. (2013). Phasic dopamine as a prediction error of intrinsic and extrinsic reinforcements driving both action acquisition and reward maximization: a simulated robotic study. *Neural Netw.* 39, 40–51. doi: 10.1016/j.neunet.2012.12.012
- Mugan, J., and Kuipers, B. (2009). "Autonomously learning an action hierarchy using a learned qualitative state representation," in *Proceedings of the 21st international joint conference on Artificial intelligence, IJCAI'09* (San Francisco, CA: Morgan Kaufmann Publishers Inc.), 1175–1180.
- Oudeyer, P.-Y., Kaplan, F., and Hafner, V. (2007a). Intrinsic motivation system for autonomous mental development. *IEEE Trans. Evol. Comput.* 11, 703–713. doi: 10.1109/TEVC.2006.890271
- Oudeyer, P.-Y., and Kaplan, F. (2007b). What is intrinsic motivation? a typology of computational approaches. *Front. Neurorobot.* 1:6. doi: 10.3389/neuro.12.006.2007
- Pouget, A., and Snyder, L. H. (2000). Computational approaches to sensorimotor transformations. *Nat. Neurosci.* 3(Suppl), 1192–1198. doi: 10.1038/81469
- Ryan, R. M., and Deci, E. L. (2000). Intrinsic and extrinsic motivations: classic definitions and new directions. *Contemp. Educ. Psychol.* 25, 54–67. doi: 10.1006/ceps.1999.1020
- Santucci, V. G., Baldassarre, G., and Mirolli, M. (2010). "Biological cumulative learning through intrinsic motivations: a simulated robotic study on the development of visually-guided reaching," in *Proceedings of the Tenth International Conference on Epigenetic Robotics*, eds B. Johansson, E. Sahin, and C. Balkenius (Lund: Lund University Cognitive Studies), 121–127.
- Santucci, V. G., Baldassarre, G., and Mirolli, M. (2012a). "Intrinsic motivation mechanisms for competence acquisition," in *IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL)*, 1–6. doi: 10.1109/DevLrn.2012.6400835
- Santucci, V. G., Baldassarre, G., and Mirolli, M. (2013a). "Cumulative learning through intrinsic reinforcements," in *Evolution, Complexity and Artificial Life* eds S. Cagnoni, M. Mirolli, and M. Villani (Berlin: Springer-Verlag).
- Santucci, V. G., Baldassarre, G., and Mirolli, M. (2013b). "Intrinsic motivation signals for driving the acquisition of multiple tasks: a simulated robotic study," in *Proceedings of the 12th International Conference on Cognitive Modelling (ICCM)* (Ottawa, ON), 1–6.
- Schembri, M., Mirolli, M., and Baldassarre, G. (2007a). "Evolving childhood's length and learning parameters in an intrinsically motivated reinforcement learning robot," in *Proceedings of the Seventh International Conference on Epigenetic Robotics*, eds L. Berthouze, G. Dhristopher, M. Littman, H. Kozima, and C. Balkenius (Lund: Lund University Cognitive Studies), 141–148.
- Schembri, M., Mirolli, M., and Baldassarre, G. (2007b). "Evolving internal reinforcers for an intrinsically motivated reinforcement-learning robot," in *Proceedings of the 6th International Conference on Development and Learning*, eds Y. Demiris, D. Mareschal, B. Scassellati, and J. Weng, (London: Imperial College), E1–E6.
- Schmidhuber, J. (1991a). "A possibility for implementing curiosity and boredom in model-building neural controllers," in *Proceedings of the International Conference on Simulation of Adaptive Behavior: From Animals to Animats*, eds J. Meyer, and S. Wilson (Cambridge, MA/London: MIT Press/Bradford Books), 222–227.
- Schmidhuber, J. (1991b). "Curious model-building control system," in *Proceedings of International Joint Conference on Neural Networks*, Vol. 2, (Singapore: IEEE), 1458–1463.
- Stout, A., and Barto, A. G. (2010). "Competence progress intrinsic motivation," in *Proceedings of the Ninth IEEE International Conference on Development and Learning*, 257–262.
- Sutton, R., and Barto, A. (1998). *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press.
- Sutton, R., Precup, D., and Singh, S. (1999). Between mdps and semi-mdps: a framework for temporal abstraction in reinforcement learning. *Artif. Intell.* 112, 181–211. doi: 10.1016/S0004-3702(99)00052-1
- Sutton, R., and Tanner, B. (2005). Temporal-difference networks. *Adv. Neural Inform. Process. Syst.* 17, 1377–1348.
- Vigorito, C., and Barto, A. (2010). Intrinsically motivated hierarchical skill learning in structured environments. *IEEE Trans. Auton. Ment. Dev.* 2, 132–143. doi: 10.1109/TAMD.2010.2050205
- White, R. (1959). Motivation reconsidered: the concept of competence. *Psychol. Rev.* 66, 297–333. doi: 10.1037/h0040934
- Wittmann, B., Daw, N., Seymour, B., and Dolan, R. (2008). Striatal activity underlies novelty-based choice in humans. *Neuron* 58, 967–973. doi: 10.1016/j.neuron.2008.04.027

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 05 July 2013; accepted: 22 October 2013; published online: 12 November 2013.

*Citation: Santucci VG, Baldassarre G and Mirolli M (2013) Which is the best intrinsic motivation signal for learning multiple skills? *Front. Neurorobot.* 7:22. doi: 10.3389/fnbot.2013.00022*

This article was submitted to the journal Frontiers in Neurorobotics.

Copyright © 2013 Santucci, Baldassarre and Mirolli. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



PowerPlay: training an increasingly general problem solver by continually searching for the simplest still unsolvable problem

Jürgen Schmidhuber*

The Swiss AI Lab IDSIA, University of Lugano, SUPSI, Lugano, Switzerland

Edited by:

Gianluca Baldassarre, Italian National Research Council, Italy

Reviewed by:

Georg Martius, Max Planck Institute for Mathematics in the Sciences, Germany

Vieri G. Santucci, Istituto di Scienze e Tecnologie della Cognizione – Consiglio Nazionale delle Ricerche, Italy

***Correspondence:**

Jürgen Schmidhuber, The Swiss AI Lab IDSIA, University of Lugano and SUPSI, Galleria 2, 6928 Manno, Switzerland

e-mail: juergen@idsia.ch

Most of computer science focuses on automatically solving given computational problems. I focus on automatically *inventing* or *discovering* problems in a way inspired by the playful behavior of animals and humans, to train a more and more general problem solver from scratch in an unsupervised fashion. Consider the infinite set of all computable descriptions of tasks with possibly computable solutions. Given a general problem-solving architecture, at any given time, the novel algorithmic framework PowerPlay (Schmidhuber, 2011) searches the space of possible *pairs* of new tasks and modifications of the current problem solver, until it finds a more powerful problem solver that provably solves all previously learned tasks plus the new one, while the unmodified predecessor does not. Newly invented tasks may require to achieve a *wow-effect* by making previously learned skills more efficient such that they require less time and space. New skills may (partially) re-use previously learned skills. The greedy search of typical PowerPlay variants uses time-optimal program search to order candidate pairs of tasks and solver modifications by their conditional computational (time and space) complexity, given the stored experience so far. The new task and its corresponding task-solving skill are those first found and validated. This biases the search toward pairs that can be described compactly and validated quickly. The computational costs of validating new tasks need not grow with task repertoire size. Standard problem solver architectures of personal computers or neural networks tend to generalize by solving numerous tasks outside the self-invented training set; PowerPlay's ongoing search for novelty keeps breaking the generalization abilities of its present solver. This is related to Gödel's sequence of increasingly powerful formal theories based on adding formerly unprovable statements to the axioms without affecting previously provable theorems. The continually increasing repertoire of problem-solving procedures can be exploited by a parallel search for solutions to additional externally posed tasks. PowerPlay may be viewed as a greedy but practical implementation of basic principles of creativity (Schmidhuber, 2006a, 2010). A first experimental analysis can be found in separate papers (Srivastava et al., 2012a,b, 2013).

Keywords: problem discovery, task invention, skill learning, general problem solver, intrinsic motivation, curiosity, creativity

1. INTRODUCTION

Given a realistic piece of computational hardware with specific resource limitations, how can one devise software for it that will solve all, or at least many, of the *a priori* unknown tasks that are in principle easily solvable on this architecture? In other words, how to build a *practical* general problem solver, given the computational restrictions? It does not need to be *universal* and *asymptotically* optimal (Levin, 1973; Hutter, 2002; Schmidhuber, 2004b, 2009) like the recent (not necessarily practically feasible) general problem solvers discussed in Section 7.2; instead it should take into account all constant architecture-specific slowdowns ignored in the asymptotic optimality notation of theoretical computer science, and be generally useful for real-world applications.

Let us draw inspiration from biology. How do initially helpless human babies become rather general problem solvers over time? Apparently by playing. For example, even in the absence of external reward or hunger they are curious about what happens if they move their eyes or fingers in particular ways, creating little experiments which lead to initially novel and surprising but eventually predictable sensory inputs, while also learning motor skills to reproduce these outcomes. (See Schmidhuber, 1991a,b, 1999, 2006a, 2010; Yi et al., 2011 and Section 7.4 for previous artificial systems of this type.) Infants continually seem to invent new tasks that become boring as soon as their solutions become known. Easy-to-learn new tasks are preferred over unsolvable or hard-to-learn tasks. Eventually the numerous skills acquired in this creative, self-supervised way may get re-used to facilitate the search

for solutions to external problems, such as finding food when hungry. While kids keep inventing new problems for themselves, they move through remarkable developmental stages (Harlow et al., 1950; Berlyne, 1954; Piaget, 1955).

Here I introduce a novel unsupervised algorithmic framework for training a computational problem solver from scratch, continually searching for the simplest (fastest to find) combination of task and corresponding task-solving skill to add to its growing repertoire, without forgetting any previous skills (Section 2), or at least without decreasing average performance on previously solved tasks (Section 6.1). New skills may (partially) re-use previously learned skills. Every new task added to the repertoire is essentially defined by the time required to invent it, to solve it, and to demonstrate that no previously learned skills got lost. The search takes into account that typical problem solvers may learn to solve tasks outside the growing self-made training set due to generalization properties of their architectures. The framework is called POWERPLAY because it continually (Ring, 1994) aims at boosting computational ability and problem-solving capacity, reminiscent of humans or human societies trying to boost their general power/capabilities/knowledge/skills in playful ways, even in the absence of externally defined goals, although the skills learned by this type of pure curiosity may later help to solve externally posed tasks.

Unlike our first implementations of curious/creative/playful agents from the 1990s (Schmidhuber, 1991a, 1999; Storck et al., 1995) (Section 7.4; compare (Barto, 2013; Dayan, 2013; Nehmzow et al., 2013; Oudeyer et al., 2013)), POWERPLAY provably (by design) does not have any problems with online learning – it cannot forget previously learned skills, automatically segmenting its life into a sequence of clearly identified tasks with explicitly recorded solutions. Unlike the task search of theoretically optimal creative agents (Schmidhuber, 2006a, 2010) (Section 7.4), POWERPLAY's task search is greedy, but at least practically feasible.

Some claim that scientists often invent appropriate problems for their methods, rather than inventing methods to solve given problems. The present paper formalizes this in a way that may be more convenient to implement than those of our previous work (Schmidhuber, 1991a, 1999, 2006a, 2010), and describes a simple practical framework for building creative artificial scientists or explorers that by design continually come up with the fastest to find, initially novel, but eventually solvable problems.

1.1. BASIC IDEAS

In traditional computer science, given some formally defined task, a search algorithm is used to search a space of solution candidates until a solution to the task is found and verified. If the task is hard the search may take long.

To automatically construct an increasingly general problem solver, let us expand the traditional search space in an unusual way, such that it includes all possible *pairs* of computable tasks with possibly computable solutions, and problem solvers. Given an old problem solver that can already solve a finite known set of previously learned tasks, a search algorithm is used to find a new pair that provably has the following properties: (1) the new task cannot be solved by the old problem solver. (2) The new task can be solved by the new problem solver (some modification of the old

one). (3) The new solver can still solve the known set of previously learned tasks.

Once such a pair is found, the cycle repeats itself. This will result in a continually growing set of known tasks solvable by an increasingly more powerful problem solver. Solutions to new tasks may (partially) re-use solutions to previously learned tasks.

Smart search (e.g., Section 4.1 and Algorithm 4.1) orders candidate pairs of the type (*task, solver*) by computational complexity, using concepts of optimal universal search (Levin, 1973; Schmidhuber, 2004b), with a bias toward pairs that can be described by few additional bits of information (given the experience so far) and that can be validated quickly.

At first glance it might seem harder to search for pairs of tasks and solvers instead of solvers only, due to the apparently larger search space. However, the additional freedom of *inventing* the tasks to be solved may actually greatly reduce the time intervals between problem solver advances, because the system may often have the option of inventing a rather simple task with an easy-to-find solution.

A new task may be about simplifying the old solver such that it can still solve all tasks learned so far, but with less computational resources such as time and storage space (e.g., Section 3.1 and Algorithm 6.1).

Since the new pair (*task, solver*) is the first one found and validated, the search automatically trades off the time-varying efforts required to either invent completely new, previously unsolvable problems, or compressing/speeding up previous solutions. Sometimes it is easier to refine or simplify known skills, sometimes to invent new skills.

On typical problem solver architectures of personal computers (PCs) or neural networks (NNs), while a limited known number of previously learned tasks has become solvable, so too has a large number of unknown, never-tested tasks (in the field of Machine Learning, this is known as *generalization*). POWERPLAY's ongoing search is continually testing (and always trying to go beyond) the generalization abilities of the most recent solver instance; some of its search time has to be spent on demonstrating that self-invented new tasks are not already solvable.

Often, however, much more time will have to be spent on making sure that a newly modified solver did not forget any of the possibly many previously learned skills. Problem solver modularization (Section 3.3, especially 3.3.2) may greatly reduce this time though, making POWERPLAY prefer pairs whose validation does not require the re-testing of too many previously learned skills, thus decomposing at least part of the search space into somewhat independent regions, realizing *divide and conquer* strategies as by-products of its built-in drive to invent and validate novel tasks/skills as quickly as possible.

A biologically inspired hope is that as the problem solver is becoming more and more general, it will find it easier and easier to solve externally posed tasks (Section 5), just like growing infants often seem to re-use their playfully acquired skills to solve teacher-given problems.

1.2. OUTLINE OF REMAINDER

Section 2 will introduce basic notation and Variant 1 of the algorithmic framework POWERPLAY, which invokes the

essential procedures **TASK INVENTION**, **SOLVER MODIFICATION**, and **CORRECTNESS DEMONSTRATION**. Section 3 will discuss details of these procedures.

More detailed instantiations of PowerPlay will be described in Section 4.3 (an evolutionary method, Algorithm 4.3) and Section 4.1 (an asymptotically optimal program search method, Algorithm 4.1).

As mentioned above, the skills acquired to solve self-generated tasks may later greatly facilitate solutions to externally posed tasks, just like the numerous motor skills learned by babies during curious exploration of its world often can be re-used later to maximize external reward. Sections 5 and 6.1 will discuss variants of the framework (e.g., Algorithm 6.1) in which some of the tasks can be defined externally.

Section 6.1 will also describe a natural variant of the framework that explicitly penalizes solution costs (including time and space complexity), and allows for forgetting aspects of previous solutions, provided the average performance on previously solved tasks does not decrease.

Section 7 will point to illustrative experiments (Section 7.8) described in separate papers (Srivastava et al., 2012b, 2013), and discuss relations to previous work.

2. NOTATION AND ALGORITHMIC FRAMEWORK POWERPLAY (VARIANT II)

B^* denotes the set of finite sequences or bitstrings over the binary alphabet $B = \{0, 1\}$, λ the empty string, x, y, z, p, q, r, u strings in B^* , \mathbb{N} the natural numbers, \mathbb{R} the real numbers, $\epsilon \in \mathbb{R}$ a positive constant, m, n, n_0, k, i, j, l non-negative integers, $L(x)$ the number of bits in x (where $L(\lambda) = 0$), f, g functions mapping integers to integers. We write $f(n) = O(g(n))$ if there exist positive c, n_0 such that $f(n) \leq cg(n)$ for all $n > n_0$.

The computational architecture of the problem solver may be a deterministic universal computer, or a more limited device such as a finite state automaton or a feedforward neural network (NN) (Bishop, 2006). All such problem solvers can be uniquely encoded (Gödel, 1931) or implemented on universal computers (Church, 1936; Post, 1936; Turing, 1936) such as universal Turing Machines (TM). Therefore, without loss of generality, the remainder of this paper assumes a fixed universal reference computer whose input programs and outputs are elements of B^* . A user-defined subset $\mathcal{S} \subset B^*$ defines the set of possible problem solvers. For example, if the problem solver's architecture is itself a binary universal TM or a standard computer, then \mathcal{S} represents its set of possible programs, or a limited subset thereof – compare Sections 3.2 and 4.1. If it is a feedforward NN, then \mathcal{S} could be a highly restricted subset of programs encoding the NN's possible topologies and weights (floating point numbers) – compare Section 7.8 and the original SLIM NN paper (Schmidhuber, 2012).

In what follows, for convenience I will often identify bitstrings in B^* with things they encode, such as integers, real-valued vectors, weight matrices, or programs – the context will always make clear what is meant.

The problem solver's initial program is called s_0 . There is a set of possible task descriptions $\mathcal{T} \subset B^*$. \mathcal{T} may be the infinite set of *all* possible computable descriptions of tasks with possibly computable solutions, or just a small subset thereof. For example, a

simple task may require the solver to answer a particular input pattern with a particular output pattern (more formal details on pattern recognition tasks are given in Section 3.1.1). Or it may require the solver to steer a robot toward a goal through a sequence of actions (more formal details on sequential decision-making tasks in unknown environments are given in Section 3.1.2). There is a particular sequence of task descriptions T_1, T_2, \dots , where each unique $T_i \in \mathcal{T}$ ($i = 1, 2, \dots$) is chosen or “invented” by a search method described below such that the solutions of T_1, T_2, \dots, T_i can be computed by s_i , the i -th instance of the program, but not by s_{i-1} ($i = 1, 2, \dots$). Each T_i consists of a unique problem identifier that can be read by s_i through some built-in input processing mechanism (e.g., input neurons of an NN (Schmidhuber, 2012)), and a unique description of a deterministic procedure for determining whether the problem has been solved. Denote $T_{\leq i} = \{T_1, \dots, T_i\}$; $T_{$

A valid task T_i ($i > 1$) may require solving at least one previously solved task T_k ($k < i$) more efficiently, by using less resources such as storage space, computation time, energy, etc., thus achieving a *Wow-effect*. See Section 3.1.

Tasks and problem solver modifications are computed and validated by elements of another appropriate set of programs $\mathcal{P} \subset B^*$. Programs $p \in \mathcal{P}$ may contain instructions for reading and executing (parts of) the code of the present problem solver and reading (parts of) a recorded history $Trace \in B^*$ of previous events that led to the present solver. The algorithmic framework (Algorithm 2) incrementally trains the problem solver by finding $p \in \mathcal{P}$ that increase the set of solvable tasks.

3. TASK INVENTION, SOLVER MODIFICATION, CORRECTNESS DEMO

A program tested by Algorithm 2 has to allocate its runtime to solve three main jobs, namely, **TASK INVENTION**, **SOLVER MODIFICATION**, **CORRECTNESS DEMONSTRATION**. Now examples of each will be listed.

3.1. IMPLEMENTING TASK INVENTION

Part of the job of $p_i \in \mathcal{P}$ is to compute $T_i \in \mathcal{T}$. This will consume some of the total computation time allocated to p_i . Two examples will be given: pattern recognition tasks are treated in Section 3.1.1; sequential decision-making tasks in Section 3.1.2.

3.1.1. Example: pattern recognition tasks

In the context of learning to recognize or analyze patterns, T_i could be a 4-tuple $(I_i, O_i, t_i, n_i) \in \mathcal{I} \times \mathcal{O} \times \mathbb{N} \times \mathbb{N}$, where $\mathcal{I}, \mathcal{O} \subset B^*$, and T_i is solved if s_i satisfies $L(s_i) < n_i$ and needs at most t_i discrete time steps to read I_i and compute O_i and halt. Here I_i itself may be a pair $(I_i^1, I_i^2) \in B^* \times B^*$, where I_i^1 is constrained to be the address of an image in a given database of patterns, and I_i^2 is a p_i -generated “query” that uniquely specifies how the image should be classified through target pattern O_i , such that the same image can be analyzed in different ways during different tasks. For example, depending on the nature of the invented task sequence, the problem solver could eventually learn that $O = 1$ if $I^2 = 1001$ (suppressing task indices) and the image addressed by I^1 contains at least one black pixel, or if $I^2 = 0111$ and the image shows a cow.

Since the definition of task T_i includes bounds n_i, t_i on computational resources, T_i may be about solving at least one T_k ($k < i$)

Algorithm 2: Algorithmic Framework PowerPlay (Variant I)

```

Initialize  $s_0$  in some way.
for  $i := 1, 2, \dots$  do
  repeat
    Let a search algorithm (examples in Section 4) create a new candidate program  $p \in \mathcal{P}$ .
    Give  $p$  limited time to do (not necessarily in this order):
      * TASK INVENTION: Let  $p$  compute a task  $T \in \mathcal{T}$ . See Section 3.1.
      * SOLVER MODIFICATION: Let  $p$  compute a value of the variable  $q \in \mathcal{S} \subset B^*$  (a candidate for  $s_i$ ) by computing a modification of  $s_{i-1}$ . See Section 3.2.
      * CORRECTNESS DEMONSTRATION: Let  $p$  try to show that  $T$  cannot be solved by  $s_{i-1}$ , but that  $T$  and all  $T_k (k < i)$  can be solved by  $q$ . See Section 3.3.
    until CORRECTNESS DEMONSTRATION was successful
    Set  $p_i := p; T_i := T; s_i := q$ ; update  $Trace$ .
end for

```

more efficiently, corresponding to a *wow-effect*. This in turn may also yield more efficient solutions to other tasks $T_l (l < i, l \neq k)$. In practical applications one may insist that such efficiency gains must exceed a certain threshold $\epsilon > 0$, to avoid task series causing sequences of very minor improvements.

Note that n_i and t_i may be unnecessary in special cases such as the problem solver being a fixed topology feedforward NN (Bishop, 2006) whose input and target patterns have constant size and whose computational efforts per pattern need constant time and space resources.

Assuming sufficiently powerful \mathcal{S}, \mathcal{P} , in the beginning, trivial tasks such as simply copying I_i^2 onto O_i may be interesting in the sense that POWERPLAY can still validate and accept them, but they will become boring (inadmissible) as soon as they are solvable by solutions to previous tasks that generalize to new tasks.

3.1.2. Example: general decision-making tasks in dynamic environments

In the more general context of general problem solving/sequential decision making/reinforcement learning/reward optimization (Newell and Simon, 1963; Kaelbling et al., 1996; Sutton and Barto, 1998) in unknown environments, there may be a set $\mathcal{I} \subset B^*$ of possible task identification patterns and a set $\mathcal{J} \subset B^*$ of programs that test properties of bitstrings. T_i could then encode a 4-tuple $(I_i, J_i, t_i, n_i) \in \mathcal{I} \times \mathcal{J} \times \mathbb{N} \times \mathbb{N}$ of finite bitstrings with the following interpretation: s_i must satisfy $L(s_i) < n_i$ and may spend at most t_i discrete time steps on first reading I_i and then interacting with an environment through a sequence of perceptions and actions, to achieve some computable goal defined by J_i .

More precisely, while T_i is being solved within t_i time steps, at any given time $1 \leq t \leq t_i$, the internal state of the problem solver at time t is denoted $u_i(t) \in B^*$; its initial default value is $u_i(0)$. For example, $u_i(t)$ may encode the current contents of the internal tape of a TM, or of certain addresses in the dynamic storage area of a PC, or the present activations of an LSTM recurrent NN (Hochreiter and Schmidhuber, 1997). At time t , s_i can spend a constant number of elementary computational instructions to copy the task description T_i or the present environmental input $x_i(t) \in B^*$ and a reward signal $r_i(t) \in B^*$ (interpreted as a real number) into parts of $u_i(t)$, then update other parts of $u_i(t)$ (a function

of $u_i(t-1)$) and compute action $y_i(t) \in B^*$ encoded as a part of $u_i(t)$. $y_i(t)$ may affect the environment, and thus future inputs.

If \mathcal{P} allows for programs that can *dynamically acquire additional physical computational resources* such as additional CPUs and storage, then the above constant number of elementary computational instructions should be replaced by a constant amount of real time, to be measured by a reliable physical clock.

The sequence of 4-tuples $(x_i(t), r_i(t), u_i(t), y_i(t)) (t = 1, \dots, t_i)$ gets recorded by the so-called trace $Trace_i \in B^*$. If at the end of the interaction a desirable computable property $J_i(Trace_i)$ (computed by applying program J_i to $Trace_i$) is satisfied, then by definition the task is solved. The set \mathcal{J} of possible J_i may represent an infinite set of all computable tasks with solutions computable by the given hardware. For practical reasons, however, the set \mathcal{J} of possible J_i may also be restricted to bit sequences encoding just a few possible goals. For example, J_i may only encode goals of the form: a robot arm steered by program or “*policy*” s_i has reached a certain target (a desired final observation $x_i(t_i)$ recorded in $Trace_i$) without measurably bumping into an obstacle along the way, that is, there were no negative rewards, that is, $r_i(\tau) \geq 0$ for $\tau = 1 \dots t_i$.

If the environment is deterministic, e.g., a digital physics simulation of a robot, then its current state can be encoded as part of $u(t)$, and it is straight-forward for CORRECTNESS DEMONSTRATION to test whether some s_i still can solve a previously solved task $T_j (j < i)$. However, what if the environment is only partially observable, like the real world, and non-stationary, changing in unknown ways? Then CORRECTNESS DEMONSTRATION must check whether s_i still produces the same action sequence in response to the input sequence recorded in $Trace_j$ (often this replay-based test will actually be computationally cheaper than a test involving the environment). Achieving the same goal in a changed environment must be considered a different task, even if the changes are just due to noise on the environmental inputs. (Sure, in the real world $s_j (j > i)$ might actually achieve J_i faster than s_i , given the description of T_i , but CORRECTNESS DEMONSTRATION in general cannot know whether this acceleration was due to plain luck – it must stick to reproducing $Trace_j$ to make sure it did not forget anything.)

See Section 6.2, however, for a less strict POWERPLAY variant whose CORRECTNESS DEMONSTRATION directly interacts with the real world to collect sufficient problem-solving statistics through

repeated trials, making certain assumptions about the probabilistic nature of the environment, and the repeatability of experiments.

3.2. IMPLEMENTING SOLVER MODIFICATION

Part of the job of $p_i \in \mathcal{P}$ is also to compute s_i , possibly profiting from having access to s_{i-1} , because only few changes of s_{i-1} may be necessary to come up with an s_i that goes beyond s_{i-1} . For example, if the problem solver is a standard PC, then just a few bits of the previous software s_{i-1} may need to be changed.

For practical reasons, the set \mathcal{S} of possible s_i may be greatly restricted to bit sequences encoding programs that obey the syntax of a standard programming language such as LISP or Java. In turn, the programming language describing \mathcal{P} should be greatly restricted such that any $p_i \in \mathcal{P}$ can only produce syntactically correct s_i .

If the problem solver is a feedforward NN with pre-wired, unmodifiable topology, then \mathcal{S} will be restricted to those bit sequences encoding valid weight matrices, s_i will encode its i -th weight matrix, and \mathcal{P} will be restricted to those $p \in \mathcal{P}$ that can produce some $s_i \in \mathcal{S}$. Depending on the user-defined programming language, p_i may invoke complex pre-wired subprograms (e.g., well-known learning algorithms) as primitive instructions – compare separate experimental analysis (Srivastava et al., 2012b, 2013).

In general, p itself determines how much time to spend on SOLVER MODIFICATION – enough time must be left for TASK INVENTION and CORRECTNESS DEMONSTRATION.

3.3. IMPLEMENTING CORRECTNESS DEMONSTRATION

Correctness demonstration may be the most time-consuming obligation of p_i . At first glance it may seem that as the sequence T_1, T_2, \dots is growing, more and more time will be needed to show that s_i but not s_{i-1} can solve T_1, T_2, \dots, T_i , because one naive way of ensuring correctness is to re-test s_i on all previously solved tasks. Theoretically more efficient ways are considered next.

3.3.1. Most general: proof search

The most general way of demonstrating correctness is to encode (in read-only storage) an axiomatic system \mathcal{A} that formally describes computational properties of the problem solver and possible s_i , and to allow p_i to search the space of possible proofs derivable from \mathcal{A} , using a proof searcher subroutine that systematically generates proofs until it finds a theorem stating that s_i but not s_{i-1} solves T_1, T_2, \dots, T_i (proof search may achieve this efficiently without explicitly re-testing s_i on T_1, T_2, \dots, T_i). This could be done like in the Gödel Machine (Schmidhuber, 2009) (Section 7.2), which uses an online extension of *Universal Search* (Levin, 1973) to systematically test *proof techniques*: proof-generating programs that may invoke special instructions for generating axioms and applying inference rules to prolong an initially empty $\text{proof} \in B^*$ by theorems, which are either axioms or inferred from previous theorems through rules such as *modus ponens* combined with *unification*, e.g., (Fitting, 1996). \mathcal{P} can be easily limited to programs generating only syntactically correct proofs (Schmidhuber, 2009). \mathcal{A} has to subsume axioms describing how any instruction invoked by some $s \in \mathcal{S}$ will change the state u of the problem solver from one step to the next (such that proof techniques can reason about the effects of any s_i). Other axioms

encode knowledge about arithmetics etc (such that proof techniques can reason about spatial and temporal resources consumed by s_i).

In what follows, CORRECTNESS DEMONSTRATIONS will be discussed that are less general but sometimes more convenient to implement.

3.3.2. Keeping track which components of the solver affect which tasks

Often it is possible to partition $s \in \mathcal{S}$ into components, such as individual bits of the software of a PC, or weights of a NN. Here the k -th component of s is denoted s^k . For each k ($k = 1, 2, \dots$) a variable list $L^k = (T_1^k, T_2^k, \dots)$ is introduced. Its initial value before the start of PowerPlay is L_0^k , an empty list. Whenever p_i found s_i and T_i at the end of CORRECTNESS DEMONSTRATION, each L^k is updated as follows: its new value L_i^k is obtained by appending to L_{i-1}^k those $T_j \notin L_{i-1}^k$ ($j = 1, \dots, i$) whose current (possibly revised) solutions now need s^k at least once during the solution-computing process, and deleting those T_j whose current solutions do not use s^k any more.

POWERPLAY's CORRECTNESS DEMONSTRATION thus has to test only tasks in the union of all L_i^k . That is, if the most recent task does not require changes of many components of s , and if the changed bits do not affect many previous tasks, then CORRECTNESS DEMONSTRATION may be very efficient.

Since every new task added to the repertoire is essentially defined by the time required to invent it, to solve it, and to show that no previous tasks became unsolvable in the process, POWERPLAY is generally “motivated” to invent tasks whose validity check does not require too much computational effort. That is, POWERPLAY will often find p_i that generate s_{i-1} -modifications that don't affect too many previous tasks, thus decomposing at least part of the spaces of tasks and their solutions into more or less independent regions, realizing *divide and conquer* strategies as by-products. Compare a recent experimental analysis of this effect (Srivastava et al., 2012b, 2013).

3.3.3. Advantages of prefix code-based problem solvers

Let us restrict \mathcal{P} such that tested $p \in \mathcal{P}$ cannot change any components of s_{i-1} during SOLVER MODIFICATION, but can create a new s_i only by adding new components to s_{i-1} . (This means freezing all used components of any s_k once T_k is found.) By restricting \mathcal{S} to self-delimiting prefix codes like those generated by the Optimal Ordered Problem Solver (OOPS) (Schmidhuber, 2004b), one can now profit from a sometimes particularly efficient type of CORRECTNESS DEMONSTRATION, ensuring that differences between s_i and s_{i-1} cannot affect solutions to $T_{<i}$ under certain conditions. More precisely, to obtain s_i , half the search time is spent on trying to process T_i first by s_{i-1} , extending or prolonging s_{i-1} only when the ongoing computation requests to add new components through special instructions (Schmidhuber, 2004b) – then CORRECTNESS DEMONSTRATION has less to do as the set $T_{<i}$ is guaranteed to remain solvable, by induction. The other half of the time is spent on processing T_i by a new sub-program with new components s'_i , a part of s_i but *not* of s_{i-1} , where s'_i may read s_{i-1} or invoke parts of s_{i-1} as sub-programs to solve $T_{\leq i}$ – only then CORRECTNESS DEMONSTRATION has to test s_i

not only on T_i but also on $T_{<i}$ (see (Schmidhuber, 2004b) for details).

A simple but not very general way of doing something similar is to interleave TASK INVENTION, SOLVER MODIFICATION, CORRECTNESS DEMONSTRATION as follows: restrict all $p \in \mathcal{P}$ such that they must define $I_i = i$ as the unique task identifier I_i for T_i (see Section 3.1.2); restrict all $s \in \mathcal{S}$ such that the input of $I_i = i$ automatically invokes sub-program s'_i , a part of s_i but *not* of s_{i-1} (although s'_i may read s_{i-1} or invoke parts of s_{i-1} as sub-programs to solve T_i). Restrict J_i to a subset of acceptable computational outcomes (Section 3.1.2). Run s_i until it halts and has computed a *novel* output acceptable by J_i that is different from all outputs computed by the (halting) solutions to $T_{<i}$; this novel output becomes T_i 's goal. By induction over i , since all previously used components of s_{i-1} remain unmodified, the set $T_{<i}$ is guaranteed to remain solvable, no matter s'_i . That is, CORRECTNESS DEMONSTRATION on previous tasks becomes trivial. However, in this simple setup there is no immediate generalization across tasks like in OOPS (Schmidhuber, 2004b) and the previous paragraph: the trivial task identifier i will always first invoke some s'_i different from all s'_k ($k \neq i$), instead of allowing for solving a new task solely by previously found code.

4. IMPLEMENTATIONS OF POWERPLAY

POWERPLAY is a general framework that allows for plugging in many different search and learning algorithms. The present section will discuss some of them.

4.1. IMPLEMENTATION BASED ON OPTIMAL ORDERED PROBLEM SOLVER OOPS

The i -th problem is to find a program $p_i \in \mathcal{P}$ that creates s_i and T_i and demonstrates that s_i but not s_{i-1} can solve T_1, T_2, \dots, T_i . This yields a perfectly ordered problem sequence for a variant of the *Optimal Ordered Problem Solver* OOPS (Schmidhuber, 2004b) (Algorithm 4.1).

While a candidate program $p \in \mathcal{P}$ is executed, at any given discrete time step $t = 1, 2, \dots$, its internal state or dynamical storage U at time t is denoted $U(t) \in B^*$ (not to be confused with the solver's internal state $u(t)$ of Section 3.1.2). Its initial default value is $U(0)$. E.g., $U(t)$ could encode the current contents of the internal tape of a TM (to be modified by p), or of certain cells in the dynamic storage area of a PC.

Once p_i is found, $p_i, s_i, T_i, Trace_i$ (if applicable; see Section 3.1.2) will be saved in unmodifiable read-only storage, possibly together with other data observed during the search so far. This may greatly facilitate the search for p_k , $k > i$, since p_k may contain instructions for addressing and reading $p_j, s_j, T_j, Trace_j$ ($j = 1, \dots, k-1$) and for copying the read code into modifiable storage U , where p_k may further edit the code, and execute the result, which may be a useful subprogram (Schmidhuber, 2004b).

Define a probability distribution $P(p)$ on \mathcal{P} to represent the searcher's initial bias (more likely programs p will be tested earlier (Levin, 1973)). P could be based on program length, e.g., $P(p) = 2^{-L(p)}$, or on a probabilistic syntax diagram (Schmidhuber, 2004a,b). See Algorithm 4.1.

OOPS keeps doubling the time limit until there is sufficient runtime for a sufficiently likely program to compute a novel,

previously unsolvable task, plus its solver, which provably does not forget previous solutions. OOPS allocates time to programs according to an asymptotically optimal universal search method (Levin, 1973) for problems with easily verifiable solutions, that is, solutions whose validity can be quickly tested. Given some problem class, if some unknown optimal program p requires $f(k)$ steps to solve a problem instance of size k and demonstrate the correctness of the result, then this search method will need at most $O(f(k)/P(p)) = O(f(k))$ steps – the constant factor $1/P(p)$ may be large but does not depend on k . Since OOPS may re-use previously generated solutions and solution-computing programs, however, it may be possible to greatly reduce the constant factor associated with plain universal search (Schmidhuber, 2004b).

The big difference to previous implementations of OOPS is that POWERPLAY has the additional freedom to define its own tasks. As always, every new task added to the repertoire is essentially defined by the time required to invent it, to solve it, and to demonstrate that no previously learned skills got lost.

4.1.1. Building on existing OOPS source code

Existing OOPS source code (Schmidhuber, 2004a) uses a FORTH-like universal programming language to define \mathcal{P} . It already contains a framework for testing new code on previously solved tasks, and for efficiently undoing all U -modifications of each tested program. The source code requires few changes to implement the additional task search described above.

4.1.2. Alternative problem solvers based on recurrent neural networks

Recurrent NNs (RNNs, e.g., (Robinson and Fallside, 1987; Werbos, 1988; Schmidhuber, 1992a; Williams and Zipser, 1994; Hochreiter and Schmidhuber, 1997)) are general computers that allow for both sequential and parallel computations, unlike the strictly sequential FORTH-like language of Section 4.1.1. They can compute any function computable by a standard PC (Schmidhuber, 1990). The original POWERPLAY report (Schmidhuber, 2011) used a fully connected RNN called RNN1 to define \mathcal{S} , where w^{lk} is the real-valued weight on the directed connection between the l -th and k -th neuron. To program RNN1 means to set the weight matrix $s = \langle w^{lk} \rangle$. Given enough neurons with appropriate activation functions and an appropriate $\langle w^{lk} \rangle$, Algorithm 4.1 can be used to train s . \mathcal{P} may itself be the set of weight matrices of a separate RNN called RNN2, computing tasks for RNN1, and modifications of RNN1, using techniques for network-modifying networks as described in previous work (Schmidhuber, 1992b, 1993a,b).

In first experiments (Srivastava et al., 2012b, 2013), a particularly suited NN called a self-delimiting NN or SLIM NN (Schmidhuber, 2012) is used. During program execution or activation spreading in the SLIM NN, lists are used to trace only those neurons and connections used at least once. This also allows for efficient resets of large NNs which may use only a small fraction of their weights per task. Unlike standard RNNs, SLIM NNs are easily combined with techniques of asymptotically optimal program search (Levin, 1973; Schmidhuber et al., 1997; Schmidhuber, 2003, 2004b) (Section 4.1). To address overfitting, instead of depending on pre-wired regularizers and hyper-parameters (Bishop, 2006),

Algorithm 4.1: Implementing PowerPlay with Procedure OOPS (Schmidhuber, 2004b)

(see text for details) - initialize s_0 and u (internal dynamic storage for $s \in \mathcal{S}$) and U (internal dynamic storage for $p \in \mathcal{P}$), where each possible p is a sequence of subprograms p', p'', p''' .

for $i := 1, 2, \dots$ **do**

set variable time limit $t_{lim} := 1$;
let the variable set H be empty;
set Boolean variable DONE := FALSE
repeat

if H is empty **then**

set $t_{lim} := 2t_{lim}$; $H := \{p \in \mathcal{P}: P(p)t_{lim} \geq 1\}$

else

choose and remove some p from H

while not DONE and less than $P(p)t_{lim}$ time was spent on p **do**

execute the next time step of the following computation:

1. Let p' compute some task $T \in \mathcal{T}$ and halt.
2. Let p'' compute $q \in \mathcal{S}$ by modifying a copy of s_{i-1} , and halt.
3. Let p''' try to show that q but not s_{i-1} can solve $T_1, T_2, \dots, T_{i-1}, T$.
If p''' was successful set DONE:=TRUE.

end while

Undo all modifications of u and U due to p . This does not cost more time than executing p in the while loop above (Schmidhuber, 2004b).

end if

until DONE

set $p_i := p$; $T_i := T$; $s_i := q$;

add a unique encoding of the 5-tuple $(i, p_i, s_i, T_i, Trace_i)$ to read-only storage
readable by programs to be tested in the future.

end for

SLIM NNs can in principle learn to select by themselves their own runtime and their own numbers of free parameters, becoming fast and *slim* when necessary. Efficient SLIM NN learning algorithms (LAs) track which weights are used for which tasks (Section 3.3.2), to greatly speed up performance evaluations in response to limited weight changes. LAs may penalize the task-specific total length of connections used by SLIM NNs implemented on the 3-dimensional brain-like multi-processor hardware to be expected in the future. This encourages SLIM NNs to solve many sub-tasks by subsets of neurons that are physically close (Schmidhuber, 2012).

4.2. ADAPTING THE PROBABILITY DISTRIBUTION ON PROGRAMS

A straight-forward extension of the above works as follows: whenever a new p_i is found, P is updated to make either only p_i or all p_1, p_2, \dots, p_i more likely. Simple ways of doing this are described in previous work (Schmidhuber et al., 1997). This may be justified to the extent that future successful programs turn out to be similar to previous ones.

4.3. IMPLEMENTATION BASED ON STOCHASTIC OR EVOLUTIONARY SEARCH

A possibly simpler but less general approach is to use an evolutionary algorithm to produce an s -modifying and task-generating program p as requested by POWERPLAY, according to Algorithm 4.3, which refers to the recurrent net problem solver of Section 4.1.2.

5. ADDING EXTERNAL TASKS

The growing repertoire of the problem solver may facilitate learning of solutions to externally posed tasks. For example, one may modify POWERPLAY such that for certain i , T_i is defined externally, instead of being invented by the system itself. In general, the resulting s_i will contain an externally inserted bias in form of code that will make some future self-generated tasks easier to find than others. It should be possible to push the system in a human-understandable or otherwise useful direction by regularly inserting appropriate external goals. See Algorithm 6.1.

Another way of exploiting the growing repertoire is to simply copy s_i for some I and use it as a starting point for a search for a solution to an externally posed task T , *without* insisting that the modified s_i also can solve T_1, T_2, \dots, T_i . This may be much faster than trying to solve T from scratch, to the extent the solutions to self-generated tasks reflect general knowledge (code) re-usable for T .

In general, however, it will be possible to design external tasks whose solutions do *not* profit from those of self-generated tasks – the latter even may turn out to slow down the search.

On the other hand, in the real world the benefits of curious exploration seem obvious. One should analyze theoretically and experimentally under which conditions the creation of self-generated tasks can accelerate the solution to externally generated tasks – see (Schmidhuber, 1991a, 1999, 2002; Storck et al., 1995; Cuccu et al., 2011; Luciw et al., 2011; Schaul et al., 2011; Yi et al., 2011) for previous simple experimental studies in this vein.

Algorithm 4.3: PowerPlay for RNNs Using Stochastic or Evolutionary Search

Randomly initialize RNN1's variable weight matrix $\langle w^{lk} \rangle$ and use the result as s_0 (see Section 4.1.2)

for $i := 1, 2, \dots$ **do**

- set Boolean variable DONE = FALSE
- repeat**

 - use a black box optimization algorithm BBOA (many are possible (Rechenberg, 1971; Gomez et al., 2008; Wierstra et al., 2008; Sehnke et al., 2010)) with adaptive parameter vector θ to create some $T \in \mathcal{T}$ (to define the task input to RNN1; see Section 3.1) and a modification of s_{i-1} , the current $\langle w^{lk} \rangle$ of RNN1, thus obtaining a new candidate $q \in \mathcal{S}$
 - if** q but not s_{i-1} can solve T and all $T_k (k < i)$ (see Sections 3.3, 3.3.2) **then**

 - set DONE = TRUE

 - end if**

- until** DONE
- set $s_i := q$; $\langle w^{lk} \rangle := q$; $T_i := T$; (also store $Trace_i$ if applicable, see Section 3.1.2). Use the information stored so far to adapt the parameters θ of the BBOA, e.g., by gradient-based search (Wierstra et al., 2008; Sehnke et al., 2010), or according to the principles of evolutionary computation (Rechenberg, 1971; Gomez et al., 2008; Wierstra et al., 2008).

end for

5.1. SELF-REFERENCE THROUGH NOVEL TASK SEARCH AS AN EXTERNAL TASK

POWERPLAY's i -th goal is to find a $p_i \in \mathcal{P}$ that creates T_i and s_i (a modification of s_{i-1}) and shows that s_i but not s_{i-1} can solve $T_{\leq i}$. As s itself is becoming a more and more general problem solver, s may help in many ways to achieve such goals in self-referential fashion. For example, the old solver s_{i-1} may be able to read a unique formal description (provided by p_i) of POWERPLAY's i -th goal, viewing it as an external task, and produce an output unambiguously describing a candidate for (T_i, s_i) . If s has a theorem prover component (Section 3.3.1), s_{i-1} might even output a full proof of (T_i, s_i) 's validity; alternatively p_i could just use the possibly suboptimal suggestions of s_{i-1} to narrow down and speed up the search. This is one of the reasons why Section 2 already mentioned that programs $p \in \mathcal{P}$ should contain instructions for reading (and running) the code of the present problem solver.

6. SOFTENING TASK ACCEPTANCE CRITERIA OF POWERPLAY

The POWERPLAY variants above insist that s may not solve new tasks at the expense of forgetting to solve any previously solved task within its previously established time and space bounds. For example, let us consider the sequential decision-making tasks from Section 3.1.2. Suppose the problem solver can already solve task $T_k = (I_k, J_k, t_k, n_k) \in \mathcal{I} \times \mathcal{J} \times \mathbb{N} \times \mathbb{N}$. A very similar but admissible new task $T_i = (I_k, J_k, t_i, n_k)$, ($i > k$), would be to solve T_k substantially faster: $t_i < t_k - \epsilon$, as long as T_i is not already solvable by s_{i-1} , and no solution to some $T_l (l < i)$ is forgotten in the process.

Here I discuss variants of POWERPLAY that soften the acceptance criteria for new tasks in various ways, for example, by allowing some of the computations of solutions to previous non-external (Section 5) tasks to slow down by a certain amount of time, provided the sum of their runtimes does not increase. This also permits the system to invent new previously unsolved tasks at the expense of slightly increasing time bounds for certain already solved non-external tasks, but without decreasing the average performance on the latter. Of course, POWERPLAY has to be modified accordingly, updating average runtime bounds when necessary.

Alternatively, one may allow for trading off space and time constraints in reasonable ways, e.g., in the style of asymptotically optimal *Universal Search* (Levin, 1973), which essentially trades one bit of additional space complexity for a runtime speedup factor of 2.

6.1. POWERPLAY VARIANT II: EXPLICITLY PENALIZING TIME AND SPACE COMPLEXITY

Let us remove time and space bounds from the task definitions of Section 3.1.2, since the modified cost-based POWERPLAY framework below (Algorithm 6.1) will handle computational costs (such as time and space complexity of solutions) more directly. In the present section, T_i encodes a tuple $(I_i, J_i) \in \mathcal{I} \times \mathcal{J}$ with interpretation: s_i must first read I_i and then interact with an environment through a sequence of perceptions and actions, to achieve some computable goal defined by J_i within a certain maximal time interval t_{max} (a positive constant). Let $t'_s(T)$ be t_{max} if s cannot solve task T , otherwise it is the time needed to solve T by s . Let $l'_s(T)$ be the positive constant l_{max} if s cannot solve T , otherwise it is the number of components of s needed to solve task T by s . The non-negative real-valued reward $r(T)$ for solving T is a positive constant r_{new} for self-defined previously unsolvable T , or user-defined if T is an external task solved by s (Section 5). The real-valued cost $Cost(s, TSET)$ of solving all tasks in a task set $TSET$ through s is a real-valued function of: all $l'_s(T)$, $t'_s(T)$ (for all $T \in TSET$), $L(s)$, and $\sum_{T \in TSET} r(T)$. For example, the cost function $Cost(s, TSET) = L(s) + \alpha \sum_{T \in TSET} [t'_s(T) - r(T)]$ encourages compact and fast solvers solving many different tasks with the same components of s , where the real-valued positive parameter α weighs space costs against time costs, and r_{new} should exceed t_{max} to encourage solutions of novel self-generated tasks, whose cost contributions should be below zero (alternative cost definitions could also take into account energy consumption etc.).

Let us keep an analog of the remaining notation of Section 3.1.2, such as $u_i(t)$, $x_i(t)$, $r_i(t)$, $y_i(t)$, $Trace_i$, $J_i(Trace_i)$. As always, if the environment is unknown and possibly changing over time, to test performance of a new solver s on a previous task T_k , only $Trace_k$ is necessary – see Section 3.1.2. As always, let $T_{\leq i}$ denote the

set containing all tasks T_1, \dots, T_i (note that if $T_i = T_k$ for some $k < i$ then it will appear only once in $T_{\leq i}$), and let $\epsilon > 0$ again define what is acceptable progress:

By Algorithm 6.1, s_i may forget certain abilities of s_{i-1} , provided that the overall performance as measured by $\text{Cost}(s_i, T_{\leq i})$ has improved, either because a new task became solvable, or previous tasks became solvable more efficiently.

Following Section 3.3, CORRECTNESS DEMONSTRATION can often be facilitated, for example, by tracking which components of s_i are used for solving which tasks (Section 3.3.2).

To further refine this approach, consider that in phase i , the list L_i^k (defined in Section 3.3.2) contains all previously learned tasks whose solutions depend on s^k . This can be used to determine the current value $\text{Val}(s_i^k)$ of some component s^k of s_i : $\text{Val}(s_i^k) = -\sum_{T \in L_i^k} \text{Cost}(s_i, T_{\leq i})$. It is a simple exercise to invent POWERPLAY variants that do not forget valuable components as easily as less valuable ones.

The implementations of Sections 4.1 and 4.3 are easily adapted to the cost-based POWERPLAY framework. Compare separate papers (Srivastava et al., 2012b, 2013).

6.2. PROBABILISTIC POWERPLAY VARIANTS

Section 3.1.2 pointed out that in partially observable and/or non-stationary unknown environments CORRECTNESS DEMONSTRATION must use Trace_k to check whether a new s_i still knows how to solve an earlier task $T_k (k < i)$. A less strict variant of POWERPLAY, however, will simply make certain assumptions about the probabilistic nature of the environment and the repeatability of trials, assuming that a limited fixed number of interactions with the real world are sufficient to estimate the costs c_i^* , c_i in Algorithm 6.1.

Another probabilistic way of softening POWERPLAY is to add new tasks without proof that s won't forget solutions to previous tasks, provided CORRECTNESS DEMONSTRATION can at least show that the probability of forgetting any previous solution is below some real-valued positive constant threshold.

7. DISCUSSION

Here I briefly mention illustrative experiments described in detail elsewhere (Srivastava et al., 2012b, 2013) and discuss certain aspects and limitations of POWERPLAY. I also discuss related research, in particular, why the present work is of interest despite the recent advent of theoretically optimal universal problem solvers (Section 7.2), and how it can be viewed as a greedy but feasible and sound implementation of the formal theory of creativity (Section 7.4).

7.1. OUTGROWING TRIVIAL TASKS – COMPRESSING PREVIOUS SOLUTIONS

What prevents POWERPLAY from inventing trivial tasks forever by extreme modularization, simply allocating a previously unused solver part to each new task, which thus becomes rather quickly verifiable, as its solution does not affect solutions to previous tasks (Section 3.3.3)? On realistic but general architectures such as PCs and RNNs, at least once the upper storage size limit of s is reached, POWERPLAY will start “compressing” previous solutions, making s

generalize in the sense that the same relatively short piece of code (some part of s) helps to solve different tasks.

With many computational architectures, this type of compression will start much earlier though, because new tasks solvable by partial reuse of earlier discovered code will often be easier to find than new tasks solvable by previously unused parts of s . This also holds for growing architectures with potentially unlimited storage space.

Compare also POWERPLAY Variant II of Section 6.1 whose tasks may explicitly require improving the average time and space complexity of previous solutions by some minimal value.

In general, however, over time the system will find it more and more difficult to invent novel tasks without forgetting previous solutions, a bit like humans find it harder and harder to learn truly novel behaviors once they are leaving behind the initial rapid exploration phase typical for babies. Experiments with various problem solver architectures (e.g., (Srivastava et al., 2012b, 2013)) help to analyze such effects in detail.

7.2. RELATION TO THEORETICALLY OPTIMAL UNIVERSAL PROBLEM SOLVERS

The new millennium brought universal problem solvers that are theoretically optimal in a certain sense. The fully self-referential (Gödel, 1931) Gödel machine (Schmidhuber, 2006b, 2009) may interact with some initially unknown, partially observable environment to maximize future expected utility or reward by solving arbitrary user-defined computational tasks. Its initial algorithm is not hardwired; it can completely rewrite itself without essential limits apart from the limits of computability, but only if a proof searcher embedded within the initial algorithm can first prove that the rewrite is useful, according to the formalized utility function taking into account the limited computational resources. Self-rewrites due to this approach can be shown to be *globally optimal*, relative to Gödel's well-known fundamental restrictions of provability (Gödel, 1931). To make sure the Gödel machine is at least *asymptotically* optimal even before the first self-rewrite, one may initialize it by Hutter's non-self-referential but *asymptotically fastest algorithm for all well-defined problems* Hsearch (Hutter, 2002), which uses a hardwired brute force proof searcher and ignores the costs of proof search. Assuming discrete input/output domains $X/Y \subset B^*$, a formal problem specification $f: X \rightarrow Y$ (say, a functional description of how integers are decomposed into their prime factors), and a particular $x \in X$ (say, an integer to be factorized), Hsearch orders all proofs of an appropriate axiomatic system by size to find programs q that for all $z \in X$ provably compute $f(z)$ within time bound $t_q(z)$. Simultaneously it spends most of its time on executing the q with the best currently proven time bound $t_q(x)$. Hsearch is as fast as the *fastest* algorithm that provably computes $f(z)$ for all $z \in X$, save for a constant factor smaller than $1 + \epsilon$ (arbitrarily small real-valued $\epsilon > 0$) and an f -specific but x -independent additive constant (Hutter, 2002). Given some problem, the Gödel machine may decide to replace Hsearch by a faster method suffering less from large constant overhead, but even if it doesn't, its performance won't be less than asymptotically optimal.

Why doesn't everybody use such universal problem solvers for all computational real-world problems? Because most real-world problems are so small that the ominous constant slowdowns

Algorithm 6.1: PowerPlay Framework (Variant II) Explicitly Handling Costs of Solving Tasks

```

Initialize  $s_0$  in some way
for  $i := 1, 2, \dots$  do
  Create new global variables  $T_i \in \mathcal{T}, s_i \in \mathcal{S}, p_i \in \mathcal{P}, c_i, c_i^* \in \mathbb{R}$  (to be fixed by the end of repeat)
  repeat
    Let a search algorithm (Section 4.1) set  $p_i$ , a new candidate program. Give  $p_i$  limited time to do:
    * TASK INVENTION: Unless the user specifies  $T_i$  (Section 5), let  $p_i$  set  $T_i$ .
    * SOLVER MODIFICATION: Let  $p_i$  set  $s_i$  by computing a modification of  $s_{i-1}$  (Section 3.2).
    * CORRECTNESS DEMONSTRATION: Let  $p_i$  compute  $c_i := \text{Cost}(s_i, T_{\leq i})$ , and  $c_i^* := \text{Cost}(s_{i-1}, T_{\leq i})$ 
  until  $c_i^* - c_i > \epsilon$  (minimal savings of costs such as time/space/etc on all tasks so far)
  Freeze/store forever  $p_i, T_i, s_i, c_i, c_i^*$ 
end for

```

(potentially relevant at least before the first self-rewrite) may be large enough to prevent the universal methods from being feasible.

POWERPLAY, on the other hand, is designed to incrementally build a *practical* more and more general problem solver that can solve numerous tasks quickly, not in the asymptotic sense, but by exploiting to the max its given particular search algorithm and computational architecture, with all its space and time limitations, including those reflected by constants ignored by the asymptotic optimality notation.

As mentioned in Section 5, however, one must now analyze under which conditions POWERPLAY's self-generated tasks can accelerate the solution to externally generated tasks (compare previous experimental studies of this type (Schmidhuber, 1991a, 1999, 2002; Storck et al., 1995)).

7.3. CONNECTION TO TRADITIONAL ACTIVE LEARNING

Traditional active learning methods (Fedorov, 1972) such as AdaBoost (Freund and Schapire, 1997) have a totally different set-up and purpose: there the user provides a set of samples to be learned, then each new classifier in a series of classifiers focuses on samples badly classified by previous classifiers. Open-ended POWERPLAY, however, (1) considers arbitrary computational problems (not necessarily classification tasks); (2) can self-invent all computational tasks; (3) takes into account all computational costs, ordering task candidates by time and space complexity, relative to the present knowledge. There is no need for a pre-defined global set of tasks that each new solver tries to solve better, instead the task set continually grows based on which task is easy to invent and validate, given what is already known.

7.4. GREEDY IMPLEMENTATION OF ASPECTS OF THE FORMAL THEORY OF CREATIVITY

The Formal Theory of Creativity (Schmidhuber, 2006a, 2010) considers agents living in initially unknown environments. At any given time, such an agent uses a reinforcement learning (RL) method (Kaelbling et al., 1996) to maximize not only expected future external reward for achieving certain goals, but also *intrinsic* reward for improving an internal model of the environmental responses to its actions, learning to better predict or compress¹

the growing history of observations influenced by its behavior, thus achieving *wow-effects*, actively learning skills to influence the input stream such that it contains previously unknown but learnable algorithmic regularities. I have argued that the theory explains essential aspects of intelligence including selective attention, curiosity, creativity, science, art, music, humor, e.g., (Schmidhuber, 2006a, 2010). Compare recent related work, e.g., (Salge et al., 2012; Barto, 2013; Dayan, 2013; Nehmzow et al., 2013; Oudeyer et al., 2013).

Like POWERPLAY, such a creative agent produces a sequence of self-generated tasks and their solutions, each task still unsolvable before learning, yet becoming solvable after learning. The costs of learning as well as the learning progress are measured, and enter the reward function. Thus, in the absence of external reward for reaching user-defined goals, at any given time the agent is motivated to invent a series of additional tasks that maximize future expected learning progress.

For example, by restricting its input stream to self-generated pairs $(I, O) \in \mathcal{I} \times \mathcal{O}$ like in Section 3.1.1, and limiting it to predict only O , given I (instead of also trying to predict future (I, O) pairs from previous ones, which the general agent would do), there will be a motivation to actively generate a sequence of (I, O) pairs such that the O are first subjectively unpredictable from their I but then become predictable with little effort, given the limitations of whatever learning algorithm is used.

Below some of POWERPLAY's apparent drawbacks are listed in light of the above, followed by certain thoughts relativizing those drawbacks.

as you are moving through your office. The natural way of greatly compressing it is to construct an internal 3D model of the office space (here I am generalizing a previous analysis of the emergence of the concept of space (Philippona et al., 2004)). The 3D model allows for re-computing the entire high-resolution video from a compact sequence of very low-dimensional eye coordinates and eye directions. (The model itself typically can be specified by far fewer bits of information than needed to store raw pixel data of a long video.) Even if the 3D model is not quite precise, only relatively few extra bits will be required to encode the observed deviations from the predictions of the model. It seems clear that the enormous compression of sensory inputs achievable through an internal 3D world model is the main reason for the latter's existence. Data compression also explains the emergence of office space-independent internal representations of *movable* objects such as pens. Many additional examples of data compression in art and science and humor can be found in previous papers (Schmidhuber, 2006a, 2010).

¹It is hard to overestimate the cognitive significance of compressing the observation history. For example, consider the video-like image sequence perceived by your brain

- Instead of maximizing future expected reward, POWERPLAY is greedy, always trying to find the simplest (easiest to find and validate) task to add to the repertoire, or the simplest way of improving the efficiency or compressibility of previous solutions, instead of looking further ahead, as a universal RL method (Schmidhuber, 2006a, 2010) would do. That is, POWERPLAY may potentially sacrifice large long-term gains for small short-term gains: the discovery of many easily solvable tasks may at least temporarily prevent it from learning to solve hard tasks.

However, on general computational architectures such as RNNs (Section 4.1.2), POWERPLAY is expected to soon run out of easy tasks that are not yet solvable, due to the architecture's limited capacity and its unavoidable generalization effects (many never-tried tasks will become solvable by solutions to the few explicitly tested T_i). Compare Section 7.1.

- The general creative agent above (Schmidhuber, 2006a, 2010) is motivated to improve performance on the entire history of previous still unsolved tasks, while POWERPLAY may discard much of this history, keeping only a selective list of previously solved tasks.

However, as the system is interacting with its environment, one could store the entire continually growing history, and make sure that \mathcal{T} always allows for defining the task of better compressing the history so far.

- POWERPLAY as in Section 2 has a binary criterion for adding knowledge (was the new task solvable without forgetting old solutions?), while the general agent (Schmidhuber, 2006a, 2010) uses a more informative information-theoretic measure.

However, the cost-based POWERPLAY framework (Algorithm 6.1) of Section 6 offers similar, more flexible options, rewarding compression or speedup of solutions to previously solved tasks.

On the other hand, drawbacks of previous implementations of formal creativity theory include:

- Some previous approximative implementations (Schmidhuber, 1991a; Storck et al., 1995) used traditional RL methods (Kaelbling et al., 1996) with theoretically unlimited look-ahead. But those are limited in many ways and not guaranteed to work well in partially observable and/or non-stationary environments where the reward function changes over time. They won't necessarily generate an optimal sequence of future tasks or experiments.
- Theoretically optimal implementations (Schmidhuber, 2006a, 2010) are currently still impractical, for reasons similar to those discussed in Section 7.2.

Hence POWERPLAY may be viewed as a greedy but feasible implementation of certain basic principles of creativity (Schmidhuber, 2006a, 2010). POWERPLAY-based systems are continually motivated to invent new tasks solvable by formerly unknown procedures, or to compress or speed up problem-solving procedures discovered earlier. Unlike previous implementations, POWERPLAY extracts from the lifelong experience history a sequence of clearly identified and separated tasks with explicitly recorded solutions.

By design it cannot suffer from online learning problems affecting its solver's performance on previously solved problems.

7.5. BEYOND ALGORITHMIC ZERO-SUM TASK-INVENTION GAMES

POWERPLAY's most closely related previous task-inventing system is the *dual brain* (Schmidhuber, 1997, 1999, 2002). There, to address the computational costs of learning, and the costs of measuring learning progress, computationally powerful encoders and problem solvers (Schmidhuber, 1997, 2002) are implemented as two very general, co-evolving, symmetric, opposing modules called the *right brain* and the *left brain*. Both are able to influence the construction of self-modifying probabilistic programs written in a universal programming language. An internal storage for temporary computational results of the programs is viewed as part of the changing environment. Each module can suggest experiments or self-invented computational tasks in the form of probabilistic algorithms to be executed, and make predictions about their effects, *betting intrinsic reward* on their outcomes. The opposing module may accept such a bet in a zero-sum game by making a contrary prediction, or reject it. In case of acceptance, the winner is determined by executing the experiment and checking its outcome; the intrinsic reward eventually gets transferred from the surprised loser to the confirmed winner. Both modules try to maximize reward using a rather general RL algorithm (the so-called success-story algorithm SSA (Schmidhuber et al., 1997)) designed for complex stochastic policies (alternative RL algorithms could be plugged in as well). Thus both modules are motivated to discover *novel* tasks exhibiting novel algorithmic patterns/compressibility (=surprising *wow-effects*), where the subjective baseline for novelty is given by what the opponent already knows about the (external or internal) world's repetitive patterns. Since the execution of any computational or physical action costs something (as it will reduce the cumulative reward per time ratio), both modules are motivated to focus on self-invented tasks that involve those parts of the dynamic world that currently make surprises and learning progress *easy*, to minimize the costs of identifying promising experiments and executing them. The system learns a partly hierarchical structure of more and more complex skills or programs necessary to solve the growing sequence of self-generated tasks, reusing previously acquired simpler skills where this is beneficial. Experimental studies exhibit several sequential developmental stages, with and without external reward (Schmidhuber, 1999, 2002).

However, the *dual brain* system (Schmidhuber, 1999, 2002) did not have a built-in guarantee that it cannot forget previously learned skills, while POWERPLAY as in Section 2 does (and the time and space complexity-based variant Algorithm 6.1 of Section 6 can forget only if this improves the average efficiency of previous solutions).

7.6. OPPOSING FORCES: IMPROVING GENERALIZATION THROUGH COMPRESSION, BREAKING GENERALIZATION THROUGH NOVELTY

Two opposing forces are at work in POWERPLAY. On the one hand, the system continually tries to improve previously learned skills, by speeding them up, and by compressing the used parameters of the problem solver, reducing its effective size. The compression drive tends to improve generalization performance, according to

the principles of *Occam's Razor* and *Minimum Description Length* (MDL) and *Minimum Message Length* (MML) (Solomonoff, 1964, 1978; Kolmogorov, 1965; Wallace and Boulton, 1968; Rissanen, 1978; Wallace and Freeman, 1987; Li and Vitányi, 1997; Hutter, 2005). On the other hand, the system also continually tries to invent new tasks that break the generalization capabilities of the present solver.

POWERPLAY's time-minimizing search for new tasks automatically manages the trade-off between these opposing forces. Sometimes it is easier (because fewer computational resources are required) to invent and solve a completely new, previously unsolvable problem. Sometimes it is easier to compress (or speed up) solutions to previously invented problems.

7.7. RELATION TO GÖDEL'S SEQUENCE OF INCREASINGLY POWERFUL AXIOMATIC SYSTEMS

In 1931, Kurt Gödel showed that for each sufficiently powerful (ω -) consistent axiomatic system there is a statement that must be true but cannot be proven from the axioms through an algorithmic theorem-proving procedure (Gödel, 1931). This unprovable statement can then be added to the axioms, to obtain a more powerful formal theory in which new formerly unprovable theorems become provable, without affecting previously provable theorems.

In a sense, POWERPLAY is doing something similar. Assume the architecture of the solver is a universal computer (Gödel, 1931; Church, 1936; Post, 1936; Turing, 1936). Its software s can be viewed as a theorem-proving procedure implementing certain enumerable axioms and computable inference rules. POWERPLAY continually tries to modify s such that the previously proven theorems remain provable within certain time bounds, and a new previously unprovable theorem becomes provable.

7.8. FIRST ILLUSTRATIVE EXPERIMENTS

First experiments with POWERPLAY were reported in separate papers (Srivastava et al., 2012b, 2013) (some experiments were also briefly mentioned in the original report (Schmidhuber, 2011)). Standard NNs as well as SLIM RNNs (Schmidhuber, 2012) were used as computational problem-solving architectures. The weights of SLIM RNNs can encode essentially arbitrary computable tasks as well as arbitrary, self-delimiting, halting or non-halting programs solving those tasks (Section 4.1.2). Such programs may affect both environment (through effectors) and internal states encoding abstractions of event sequences. For example, in the experiments a SLIM RNN learned to control a fovea that can be shifted across a visual scene. The sequences of dynamically

REFERENCES

- Barto, A. (2013). "Intrinsic motivation and reinforcement learning," in *Intrinsically Motivated Learning in Natural and Artificial Systems*, eds G. Baldassarre, and M. Mirolli (Berlin: Springer), 17–47.
- Berlyne, D. E. (1954). A theory of human curiosity. *Br. J. Psychol.* 45, 180–191.
- Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.
- Church, A. (1936). An unsolvable problem of elementary number theory. *Am. J. Math.* 58, 345–363. doi:10.2307/2371045
- Cuccu, G., Luciw, M., Schmidhuber, J., and Gomez, F. (2011). "Intrinsically motivated evolutionary search for vision-based reinforcement learning," in *Proceedings of the 2011 IEEE Conference on Development and Learning and Epigenetic Robotics IEEE-ICDL-EPIROB* (IEEE).
- Dayan, P. (2013). "Exploration from generalization mediated by multiple controllers," in *Intrinsically Motivated Learning in Natural and Artificial Systems*, eds G. Baldassarre, and M. Mirolli (Berlin: Springer), 73–91.
- Fedorov, V. V. (1972). *Theory of Optimal Experiments*. Academic Press.
- Fitting, M. C. (1996). *First-Order Logic and Automated Theorem Proving*. Graduate Texts in Computer Science, 2nd Edn. Berlin: Springer-Verlag.
- Freund, Y., and Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.* 55, 119–139. doi:10.1006/jcss.1997.1504
- Gödel, K. (1931). Über formal unentscheidbare Sätze der principia mathematica und verwandter systeme I. *Monatsh. Mathematik Physik* 38, 173–198. doi:10.1007/BF01700692

changing sensory inputs from the fovea contributed to the formation of internal SLIM RNN states, that is, vectors of neural activations encoding possible goals. In open-ended fashion, our POWERPLAY-driven NNs learned to become increasingly general solvers of self-invented tasks. Sometimes they added new problem-solving procedures to the growing repertoire. Sometimes they preferred to compress/speed up previously invented skills, depending on what was computationally easiest at this point in time. The NNs also exhibited interesting developmental stages, incrementally moving from apparently simple self-invented problems to more complex ones. Furthermore, it was shown how a POWERPLAY-driven SLIM NN automatically self-modularizes (Srivastava et al., 2013), frequently re-using code for previously invented skills, keeping track which connections affect which tasks (Section 3.3.2), always trying to invent novel tasks that can be quickly validated because they do not require too many weight changes affecting too many previous tasks.

8. WORDS OF CAUTION

The behavior of POWERPLAY is determined by the nature and the limitations of \mathcal{T} , \mathcal{S} , \mathcal{P} , and its algorithm for searching \mathcal{P} . If \mathcal{T} includes all computable task descriptions, and both \mathcal{S} and \mathcal{P} allow for implementing arbitrary programs, and the search algorithm is a general method for search in program space (Section 4), then there are few limits to what POWERPLAY may do (besides the limits of computability (Gödel, 1931)).

It may not be advisable to let a general variant of POWERPLAY loose in an uncontrolled situation, e.g., on a multi-computer network on the internet, possibly with access to control of physical devices, and the potential to acquire additional computational and physical resources (Section 3.1.2) through programs executed during POWERPLAY. Unlike, say, traditional virus programs, POWERPLAY-based systems will continually change in a way hard to predict, incessantly inventing and solving novel, self-generated tasks, only driven by a desire to increase their general problem-solving capacity, perhaps a bit like many humans seek to increase their power once their basic needs are satisfied. This type of artificial curiosity/creativity, however, may conflict with human intentions on occasion. On the other hand, unchecked curiosity may sometimes also be harmful or fatal to the learning system itself (Section 5) – curiosity can kill the cat.

ACKNOWLEDGMENTS

Thanks to Mark Ring, Bas Steunebrink, Faustino Gomez, Sohrob Kazerounian, Hung Ngo, Leo Pape, Giuseppe Cuccu, and several anonymous reviewers, for useful comments.

- Gomez, F. J., Schmidhuber, J., and Miikkulainen, R. (2008). Accelerated neural evolution through cooperatively coevolved synapses. *J. Mach. Learn. Res.* 9, 937–965.
- Harlow, H. E., Harlow, M. K., and Meyer, D. R. (1950). Novelty and curiosity as determinants of exploratory behavior. *J. Exp. Psychol.* 41, 68–80.
- Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.* 9, 1735–1780. doi:10.1162/neco.1997.9.8.1735
- Hutter, M. (2002). The fastest and shortest algorithm for all well-defined problems. *Int. J. Found. Comput. Sci.* 13, 431–443. doi:10.1142/S0129054102001199 (On J. Schmidhuber's SNF grant 20-61847).
- Hutter, M. (2005). *Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability*. Berlin: Springer. (On J. Schmidhuber's SNF grant 20-61847).
- Kaelbling, L. P., Littman, M. L., and Moore, A. W. (1996). Reinforcement learning: a survey. *J. AI Res.* 4, 237–285.
- Kolmogorov, A. N. (1965). Three approaches to the quantitative definition of information. *Probl. Inform. Transm.* 1, 1–11.
- Levin, L. A. (1973). Universal sequential search problems. *Probl. Inform. Transm.* 9, 265–266.
- Li, M., and Vitányi, P. M. B. (1997). *An Introduction to Kolmogorov Complexity and its Applications*, 2nd Edn. Springer.
- Luciw, M., Graziano, V., Ring, M., and Schmidhuber, J. (2011). “Artificial curiosity with planning for autonomous perceptual and cognitive development,” in *Proceedings of the First Joint Conference on Development Learning and on Epigenetic Robotics ICDL-EPIROB*, Frankfurt.
- Nehmzow, U., Gatsoulis, Y., Kerr, E., Condell, J., Siddique, N. H., and McGinnity, T. M. (2013). “Novelty detection as an intrinsic motivation for cumulative learning robots,” in *Intrinsically Motivated Learning in Natural and Artificial Systems*, eds G. Baldassarre, and M. Mirolli (Berlin: Springer), 185–207.
- Newell, A., and Simon, H. (1963). “GPS, a program that simulates human thought,” in *Computers and Thought*, eds E. Feigenbaum, and J. Feldman (New York: McGraw-Hill), 279–293.
- Oudeyer, P.-Y., Baranes, A., and Kaplan, F. (2013). “Intrinsically motivated learning of real world sensorimotor skills with developmental constraints,” in *Intrinsically Motivated Learning in Natural and Artificial Systems*, eds G. Baldassarre, and M. Mirolli (Berlin: Springer), 303–365.
- Philipona, D., O'Regan, J. K., and Nadal, J. P. (2004). “Perception of the structure of the physical world using unknown sensors and effectors,” in *Advances in Neural Information Processing Systems*, Vol. 16 (MIT Press), 945–952.
- Piaget, J. (1955). *The Child's Construction of Reality*. London: Routledge and Kegan Paul.
- Post, E. L. (1936). Finite combinatorial processes-formulation 1. *J. Symbol. Log.* 1, 103–105. doi:10.2307/2269031
- Rechenberg, I. (1971). *Evolutionsstrategie – Optimierung Technischer Systeme nach Prinzipien der Biologischen Evolution*. Dissertation, Frommann-Holzboog, Stuttgart.
- Ring, M. B. (1994). *Continual Learning in Reinforcement Environments*. Ph.D. thesis, University of Texas at Austin, Austin, TX.
- Rissanen, J. (1978). Modeling by shortest data description. *Automatica* 14, 465–471. doi:10.1016/0005-1098(78)90005-5
- Robinson, A. J., and Fallside, F. (1987). *The Utility Driven Dynamic Error Propagation Network*. Technical Report CUED/F-INFENG/TR.1. Cambridge: Cambridge University Engineering Department.
- Salge, C., Glackin, C., and Polani, D. (2012). Approximation of empowerment in the continuous domain. *Adv. Complex Syst.* 16, 1250079. doi:10.1142/S0219525912500798
- Schaul, T., Yi, S., Wierstra, D., Gomez, F., and Schmidhuber, J. (2011). “Curiosity-driven optimization,” in *IEEE Congress on Evolutionary Computation (CEC)*, New Orleans.
- Schmidhuber, J. (1990). *Dynamische neuronale Netze und das fundamentale raumzeitliche Lernproblem*. Dissertation, Institut für Informatik, Technische Universität München, München.
- Schmidhuber, J. (1991a). “Curious model-building control systems,” in *Proceedings of the International Joint Conference on Neural Networks*, Vol. 2 (Singapore: IEEE Press), 1458–1463.
- Schmidhuber, J. (1991b). “A possibility for implementing curiosity and boredom in model-building neural controllers,” in *Proceedings of the International Conference on Simulation of Adaptive Behavior: From Animals to Animats*, eds J. A. Meyer, and S. W. Wilson (MIT Press/Bradford Books), 222–227.
- Schmidhuber, J. (1992a). A fixed-size storage $O(n^3)$ time complexity learning algorithm for fully recurrent continually running networks. *Neural Comput.* 4, 243–248. doi:10.1162/neco.1992.4.2.243
- Schmidhuber, J. (1992b). Learning to control fast-weight memories: an alternative to recurrent nets. *Neural Comput.* 4, 131–139. doi:10.1162/neco.1992.4.1.131
- Schmidhuber, J. (1993a). “On decreasing the ratio between learning complexity and number of time-varying variables in fully recurrent nets,” in *Proceedings of the International Conference on Artificial Neural Networks* (Amsterdam: Springer), 460–463.
- Schmidhuber, J. (1993b). “A self-referential weight matrix,” in *Proceedings of the International Conference on Artificial Neural Networks* (Amsterdam: Springer), 446–451.
- Schmidhuber, J. (1997). *What's Interesting?* Technical Report IDSIA-35-97. IDSIA. Available at: <ftp://ftp.idsia.ch/pub/juergen/interest.ps.gz>; extended abstract in Proceedings of the Snowbird'98, UT.
- Schmidhuber, J. (1999). “Artificial curiosity based on discovering novel algorithmic predictability through coevolution,” in *Congress on Evolutionary Computation*, eds P. Angeline, Z. Michalewicz, M. Schoenauer, X. Yao, and Z. Zalzala (IEEE Press), 1612–1618.
- Schmidhuber, J. (2002). “Exploring the predictable,” in *Advances in Evolutionary Computing*, eds A. Ghosh, and S. Tsutsui (Springer), 579–612.
- Schmidhuber, J. (2003). “Bias-optimal incremental problem solving,” in *Advances in Neural Information Processing Systems 15 (NIPS 15)*, eds S. Becker, S. Thrun, and K. Obermayer (Cambridge, MA: MIT Press), 1571–1578.
- Schmidhuber, J. (2004a). *OOPS Source Code in Crystalline Format*. Available at: <http://www.idsia.ch/~juergen/oopscode.c>
- Schmidhuber, J. (2004b). Optimal ordered problem solver. *Mach. Learn.* 54, 211–254. doi:10.1023/B:MACH.0000015880.99707.b2
- Schmidhuber, J. (2006a). Developmental robotics, optimal artificial curiosity, creativity, music, and the fine arts. *Conn. Sci.* 18, 173–187. doi:10.1080/09540090600768658
- Schmidhuber, J. (2006b). “Gödel machines: fully self-referential optimal universal self-improvers,” in *Artificial General Intelligence*, eds B. Goertzel, and C. Penncachin (Springer Verlag), 199–226. arXiv:cs.LO/0309048.
- Schmidhuber, J. (2009). Ultimate cognition à la Gödel. *Cognit. Comput.* 1, 177–193. doi:10.1007/s12559-009-9014-y
- Schmidhuber, J. (2010). Formal theory of creativity, fun, and intrinsic motivation (1990–2010). *IEEE Trans. Auton. Ment. Dev.* 2, 230–247. doi:10.1109/TAMD.2010.2056368
- Schmidhuber, J. (2011). *POWERPLAY: Training an Increasingly General Problem Solver by Continually Searching for the Simplest Still Unsolvable Problem*. Technical Report arXiv:1112.5309v1 [cs.AI].
- Schmidhuber, J. (2012). *Self-Delimiting Neural Networks*. Technical Report IDSIA-08-12, arXiv:1210.0118v1 [cs.NE], IDSIA.
- Schmidhuber, J., Zhao, J., and Wiering, M. (1997). Shifting inductive bias with success-story algorithm, adaptive Levin search, and incremental self-improvement. *Mach. Learn.* 28, 105–130. doi:10.1023/A:1007383707642
- Schnke, F., Osendorfer, C., Rückstieß, T., Graves, A., Peters, J., and Schmidhuber, J. (2010). Parameter-exploring policy gradients. *Neural Netw.* 23, 551–559. doi:10.1016/j.neunet.2009.12.004
- Solomonoff, R. J. (1964). A formal theory of inductive inference. Part I. *Inf. Control* 7, 1–22. doi:10.1016/S0019-9958(64)90131-7
- Solomonoff, R. J. (1978). Complexity-based induction systems. *IEEE Trans. Inf. Theory* IT-24, 422–432. doi:10.1109/TIT.1978.1055913
- Srivastava, R. K., Steunebrink, B. R., and Schmidhuber, J. (2012a). *First Experiments with POWERPLAY*. Technical Report arXiv:1210.8385v1 [cs.AI].
- Srivastava, R. K., Steunebrink, B. R., Stollenga, M., and Schmidhuber, J. (2012b). “Continually adding self-invented problems to the repertoire: first experiments with POWERPLAY,” in *Proceedings of the 2012 IEEE Conference on Development and Learning and Epigenetic Robotics IEEE-ICDL-EPIROB*, San Diego.
- Srivastava, R. K., Steunebrink, B. R., and Schmidhuber, J. (2013). First experiments with POWERPLAY. *Neural Netw.* 41, 130–136. doi:10.1016/j.neunet.2013.01.022
- Storck, J., Hochreiter, S., and Schmidhuber, J. (1995). “Reinforcement driven information acquisition in non-deterministic environments,” in *Proceedings of the International Conference on Evolutionary Computation*, 1995, 199–226.

- Conference on Artificial Neural Networks, Vol. 2 (Paris: EC2 & Cie), 159–164.
- Sutton, R. S., and Barto, A. G. (1998). *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press.
- Turing, A. M. (1936). On computable numbers, with an application to the Entscheidungsproblem. *Proc. Lond. Math. Soc.* 41, 230–267. (Series 2).
- Wallace, C. S., and Boulton, D. M. (1968). An information theoretic measure for classification. *Comput. J.* 11, 185–194. doi:10.1093/comjnl/11.2.185
- Wallace, C. S., and Freeman, P. R. (1987). Estimation and inference by compact coding. *J. R. Stat. Soc. B Stat. Methodol.* 49, 240–265.
- Werbos, P. J. (1988). Generalization of backpropagation with application to a recurrent gas market model. *Neural Netw.* 1, doi:10.1016/0893-6080(88)90007-X
- Wierstra, D., Schaul, T., Peters, J., and Schmidhuber, J. (2008). “Natural evolution strategies,” in *Congress of Evolutionary Computation*.
- Williams, R. J., and Zipser, D. (1994). “Gradient-based learning algorithms for recurrent networks and their computational complexity,” in *Back-Propagation: Theory, Architectures and Applications* (Hillsdale, NJ: Erlbaum).
- Yi, S., Gomez, F., and Schmidhuber, J. (2011). “Planning to be surprised: optimal Bayesian exploration in dynamic environments,” in *Proceedings of the Fourth Conference on Artificial General Intelligence (AGI)*. Mountain View, CA: Google.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 05 February 2013; accepted: 15 May 2013; published online: 07 June 2013.

Citation: Schmidhuber J (2013) PowerPlay: training an increasingly general problem solver by continually searching for the simplest still unsolvable problem. *Front. Psychol.* 4:313. doi: 10.3389/fpsyg.2013.00313

This article was submitted to Frontiers in Cognitive Science, a specialty of Frontiers in Psychology.

Copyright © 2013 Schmidhuber. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in other forums, provided the original authors and source are credited and subject to any copyright notices concerning any third-party graphics etc.



Confidence-based progress-driven self-generated goals for skill acquisition in developmental robots

Hung Ngo*, Matthew Luciw, Alexander Förster and Jürgen Schmidhuber

IDSIA, Dalle Molle Institute for Artificial Intelligence, Università della Svizzera Italiana-Scuola Universitaria Professionale della Svizzera Italiana (USI-SUPSI), Lugano, Switzerland

Edited by:

Tom Stafford, University of Sheffield, UK

Reviewed by:

Lisa Meeden, Swarthmore College, USA

Jan H. Metzen, Universität Bremen, Germany

***Correspondence:**

Hung Ngo, IDSIA, Galleria 2, 6928 Manno-Lugano, Switzerland
e-mail: hung@idsia.ch

A reinforcement learning agent that autonomously explores its environment can utilize a curiosity drive to enable continual learning of skills, in the absence of any external rewards. We formulate curiosity-driven exploration, and eventual skill acquisition, as a selective sampling problem. Each environment setting provides the agent with a stream of instances. An instance is a sensory observation that, when queried, causes an outcome that the agent is trying to predict. After an instance is observed, a query condition, derived herein, tells whether its outcome is statistically known or unknown to the agent, based on the confidence interval of an online linear classifier. Upon encountering the first unknown instance, the agent “queries” the environment to observe the outcome, which is expected to improve its confidence in the corresponding predictor. If the environment is in a setting where all instances are known, the agent generates a plan of actions to reach a new setting, where an unknown instance is likely to be encountered. The desired setting is a self-generated goal, and the plan of action, essentially a program to solve a problem, is a skill. The success of the plan depends on the quality of the agent’s predictors, which are improved as mentioned above. For validation, this method is applied to both a simulated and real Katana robot arm in its “blocks-world” environment. Results show that the proposed method generates sample-efficient curious exploration behavior, which exhibits developmental stages, continual learning, and skill acquisition, in an intrinsically-motivated playful agent.

Keywords: intrinsic motivation, artificial curiosity, continual learning, developmental robotics, online active learning, markov decision processes, AI planning, systematic exploration

1. INTRODUCTION

During our lifetimes, we continually learn, and our learning is often intrinsically motivated (Piaget, 1955; Berlyne, 1966). We do not just learn declarative knowledge, such as that exhibited by contestants appearing on the popular quiz show *Jeopardy*, but also procedural knowledge, such as how to write a Ph.D. thesis. In general, a skill is a program able to solve a limited set of problems (Schmidhuber, 1997; Srivastava et al., 2013), but the notion of a skill is often coupled with procedural knowledge, which is typically demonstrated through action. In continually learning artificial agents, skill acquisition (Newell et al., 1959; Ring, 1994; Barto et al., 2004; Konidaris, 2011; Lang, 2011; Sutton et al., 2011) is a process involving the *discovery* of new skills, learning to *reproduce* the skills reliably and efficiently, and *building upon* the acquired skills to support the acquisition of more skills. This process should never stop. An eventual goal of ours, and others, is the development of lifelong learning robot agents (Ring, 1994; Thrun and Mitchell, 1995; Ring, 1997; Sutton et al., 2011).

Traditional Markovian Reinforcement Learning (RL) (Sutton and Barto, 1998; Szepesvári, 2010) provides a formal framework that facilitates autonomous skill acquisition. In the Markov Decision Process (MDP) framework, a skill is represented as a policy that, when executed, is guaranteed to efficiently reach a particular state, which would be a “goal” state for that skill. RL involves optimizing a policy, to allow the agent to achieve the maximum expected reward.

There exist iterative *planning* methods, such as value iteration (Bellman, 1957) and policy iteration (Howard, 1960), to find an optimal policy for an MDP if a *model* of the environment is *known* to the agent; see (Mausam and Kolobov, 2012) for recent reviews. The model is the set of transition probabilities $P(s_{t+1}|s_t, a_t)$ of reaching successor state s_{t+1} , together with the associated expected immediate rewards $R(s_t, a_t)$ when the agent takes action a_t in state s_t . By selecting different goal states and creating appropriate “phantom” rewards, which are not provided by the environment, the agent could calculate a policy for a self-generated goal immediately through planning (Luciw et al., 2011; Hester and Stone, 2012; Ngo et al., 2012). An autonomous skill learner for model-based Markovian RL needs only learn a single transition model (or another type of predictive world model) and to be able to generate a different reward function for each skill.

An important issue in learning a world model is *systematic exploration*. How can an agent explore the environment to quickly and effectively learn? Early methods were based on common-sense heuristics such as “visit previously unvisited states,” or “visit states that have not been visited in a while” (Sutton, 1990). More recent methods are those based on *Artificial Curiosity* (Schmidhuber, 1991; Storck et al., 1995; Wiering and Schmidhuber, 1998; Meuleau and Bourgine, 1999; Barto et al., 2004; Şimşek and Barto, 2006; Schmidhuber, 2010; Ngo et al., 2011), which can be exploited in developmental robotics (Weng et al., 2001; Lungarella et al., 2003; Oudeyer

et al., 2007; Asada et al., 2009; Hester and Stone, 2012, Ngo et al., 2012).

Artificial curiosity uses an intrinsic reward, which is the *learning progress*, or expected improvement, of the adaptive world model [i.e., predictor/compressor of the agent's growing history of perceptions and actions (Schmidhuber, 2006)]. The expected learning progress becomes an intrinsic reward for the reinforcement learner. To maximize expected intrinsic reward accumulation, the reinforcement learner is motivated to create new experiences such that the adaptive learner makes quick progress.

We investigate an autonomous learning system that utilizes such a progress-based curiosity drive to explore its environment. This is a “pure exploration” setting, as there are no external rewards. The general framework is formulated as a selective sampling problem in which an agent samples any action in its current situation as soon as it sees that the effects of this action are statistically unknown. We present one possible implementation of the framework, using online linear classifiers (Azoury and Warmuth, 2001; Vovk, 2001; Cesa-Bianchi and Lugosi, 2006) as *predictive action models*, which essentially predict some aspects of the next state, given the current state-action features.

If no available actions have a statistically unknown outcome, the agent generates a plan of actions to reach a new setting where it expects to find such an action. The planning is implemented using approximate policy iteration, and depends on the procedural knowledge accumulated so far in the adaptive world model. The agent acquires a collection of skills through these self-generated exploration goals and the associated plans.

The framework is applied to a simulated and actual Katana robot arm manipulating blocks. Results show that our method is able to generate sample-efficient curious exploratory behavior, which exhibits developmental stages, continual learning, and skill acquisition, in an intrinsically motivated playful agent. Specifically, a desirable characteristic of a lifelong learning agent is exhibited: it should gradually move away from learned skills to focus on yet unknown but learnable skills. One particularly notable skill learned, as a by-product of its curiosity-satisfying drive, is the stable placement of a block. Another skill learned is that of stacking several blocks.

2. MATERIALS AND METHODS

In this section, we describe the setting of the learning environment, followed by introducing the selective sampling formulation (which is not environment specific). We then describe the planner and the online learning of the world model, and finally present the derivation of the query condition.

2.1. KATANA IN ITS BLOCKS-WORLD ENVIRONMENT

Our robot, a Katana arm (Neuronics, 2004), and its environment, called blocks-world, are shown in **Figure 1**. There are four different colored blocks scattered in the robot's play area. In Section 3.1 we describe a simulated version of blocks-world with eight blocks. We use the simulated version for a thorough evaluation of our method. In both versions, the agent “plays” with the blocks, through the curiosity-driven exploration framework, and learns how the world works.



FIGURE 1 | The Katana robot arm in its blocks-world environment.

In the real-world environment, detection and localization of the blocks is done with straightforward computer vision techniques. The overhead camera was calibrated using the toolbox developed by Bouguet (2009), so that the system can convert 2D image coordinates to the robot's arm-centered Cartesian coordinates. Since all the blocks have different colors, a color-based detection and pixel grouping is used for segmentation, leading to a perceptual system that reliably detects the positions and orientations (in the image coordinate system) of the visible, non-occluded blocks. The positions and orientations of occluded blocks are stored in a memory module. Since any occluded block was once a fully visible block, and the occluded block positions do not change, the memory module updating is also straightforward, requiring basic logic. The purpose of the memory module is to infer the heights of the blocks on top of occluded ones, since the overhead camera does not provide the height information.

When a block is selected for grasping, or a location selected for placement, the system converts the image coordinates to the arm-centered Cartesian coordinates. For reaching and grasping, we use the Katana's inverse kinematics module, which solves for joint angles given the desired pose (position and orientation) of the gripper, and its motion planning module.

In each environment setting, defined as a configuration of blocks, the agent first moves the gripper out of view of the camera, and takes a snapshot of the workspace below. The fundamental choice it needs to make is to decide what the most interesting block *placement location* would be. A placement location is specified by a vector including pixel-coordinates and orientation parameters in the workspace image, as well as the height, in terms of the number of blocks. After the desired placement location is decided, the agent needs to decide which block to pick up for placement. The block that is grasped could be selected via a variety of heuristics. We choose to have the robot grasp the accessible (e.g., non-occluded) block furthest away from the

desired placement location, which avoids interference with the blocks at the selected placement location. Grasping will succeed as long as the perception is accurate enough and the block is within the workspace. In the real experiments, grasping is rarely not successful. In these cases, we reset the situation (including internal values related to learning) and have the robot do it again. After grasping, the robot performs another reach, while holding a block, and places it at the desired location.

Next we will illustrate how the robot represents its world, and how this representation leads to something resembling, and which, functionally, serves as an MDP.

2.2. FOVEA AND GRAPH REPRESENTATION

The top-down camera image (640×480 pixels) is searched using a subwindow of 40×40 pixels, which we call a *fovea*. Each fovea center location represents a possible block placement location.

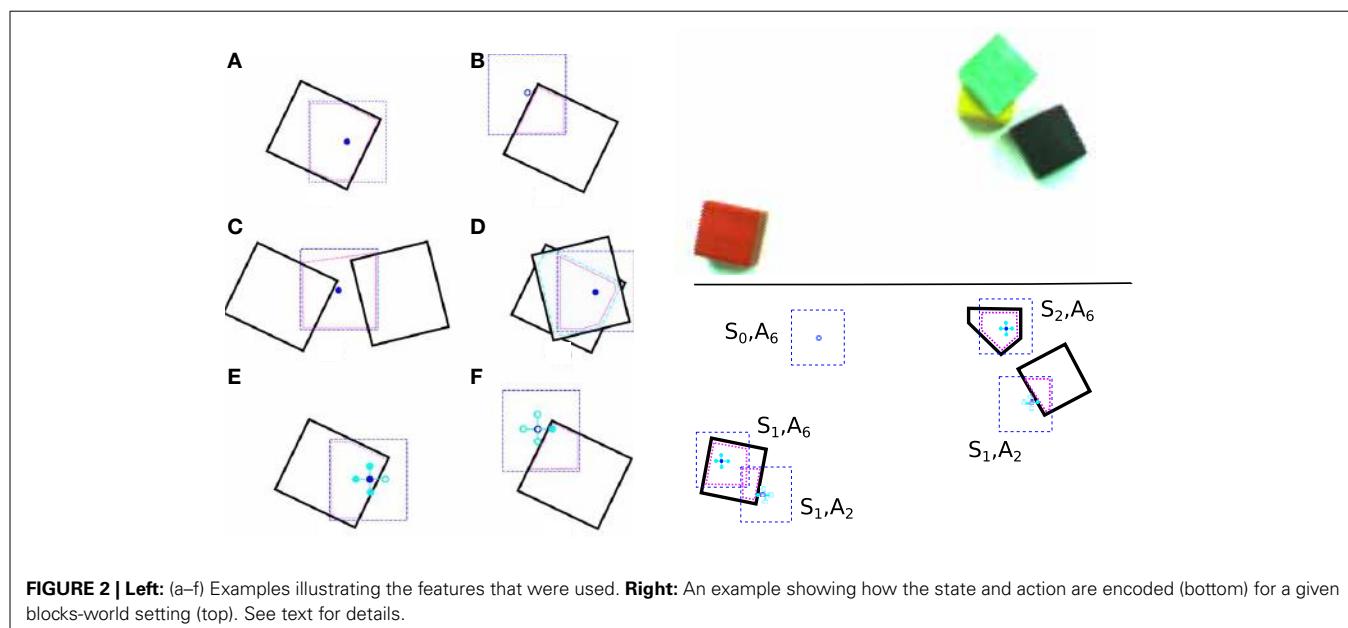
At any fovea location, the *state* s is the maximum height of a stack of blocks visible in the fovea window. The *action* a is a function of the feature vector that encapsulates the placement location relative to the blocks in any stacks below. How this feature vector is computed will be described below. Any feature vector is converted into one of six possible actions. After an action is executed, i.e., a block is picked and placed at the fovea central location, the *outcome state* s' is identified in the same way as s , with the fovea location unchanged. The resulting graph resembles a discrete MDP and serves as a basis for tractable exploration in the blocks-world environment.

In a given setting (block configuration), each fovea location maps onto a single (s, a, s') transition in a graph. But only s and a are visible before the placement experiment. The missing piece of knowledge, which the agent needs to place a block to acquire, is the outcome state s' . The fovea can be thought of as a window into a “world” where the robot can do an experiment. Yet, what the robot learns in one “world” applies to all other “worlds.” The question is: which transition is most worth sampling?

Instead of being provided a single state and having to choose an action, as in a classical RL formulation, our system is able to choose one of multiple available state-action pairs from each setting. Availability is determined from the known block positions. The agent’s estimated *global state-action value function* $Q(s, a)$ is used to identify an available state-action pair (s_t^*, a_t^*) with the highest value, constrained by availability. The agent knows the heights of all blocks in the workspace, which informs it of the possible states currently available. It also knows the fovea location that centers on each block. The desired state s_t^* is selected from the available heights in the current setting, by selecting the one with maximum state value. Next, the desired action a_t^* is selected as the one with maximum Q-value of all action pairings with s_t^* . To find a fovea location for the desired (s_t^*, a_t^*) , the agent *searches* by moving the fovea to different placement locations around the stacks of height s_t^* , until the contextual information (feature vector \mathbf{x}_t) associated with the action is matched.

The fovea search occurs in this “top-down” way, since it is computationally burdensome to extract the contextual information of state-action pairs at all fovea positions in each setting. This biased and informed search mechanism is much more efficient. As a future extension, fovea movement would be learned as well [(Whitehead and Ballard, 1990; Schmidhuber and Huber, 1991); see also recent work by Butko and Movellan (2010)].

Figure 2 (left) shows six examples to illustrate the features used. The thick black lines represent the boundaries of actual blocks. Example fovea locations are represented by the blue dashed squares. The central point of the fovea is shown as a small blue circle. The pink dotted lines show the *convex hulls* constructed from the block pixels *inside the fovea*. If the central placement point is *inside* the convex hull, the feature value is set to one, and zero otherwise. Note the case shown in (c), where the central placement point is not on top of any block at the fovea, but still within the convex hull, and so the feature is set to one. For stacks of several blocks as in (d), the *intersection* of all the



block pixels are constructed, and used to construct the convex hull.

As shown in (e) and (f), the features are calculated around the central location, which results in a five-element feature group. In our real robot implementation we use this setup. With a placement location as in (e) four bits are “on,” while in (f), only one bit is “on.” The number of bits that are on, plus one, provides the action index. For example, a fovea location with only one bit on, as described above, would correspond to action $a = A_2$ and is encoded by feature vector $\mathbf{x} = (0, 1, 0, 0, 0, 0)$. **Figure 2** (right) shows an illustration of states and actions at different fovea locations for a particular block configuration. In the lower subfigure, we see the state-action representation underneath a few sampled fovea locations. This representation allows for generalization: the same state-action (S_1, A_2) can be accessed at both the red block (to the lower right) and the black block.

We note in passing that this Katana and blocks-world environment is simplified to become functionally discrete, but the method we use for learning, approximate policy iteration, is not tabular (as the name suggests), nor is the way we use linear basis functions to convert each observation to a feature vector. Our general framework, which will be described next, does not require a tabular environment. Furthermore, the subsystem relevant to a “placement experiment,” i.e., the blocks in the stack right below the fovea, is an MDP according to the formulated graph we use. The approach of considering only relevant features in learning and planning makes the learning, and particularly the planning process, more efficient, as well as tractable¹.

2.3. SELECTIVE SAMPLING FORMULATION

Consider an online learning scenario where a learner \mathcal{L} interacts with nature \mathcal{N} (its environment) in rounds. At each round i , nature presents a *setting* \mathcal{S}_i . A setting may refer to a single state, or a set of subsystem states (as in our Katana blocks-world environment). Within *each* setting, the learner will observe a sequence of instances $\mathbf{x}_t \in \mathbb{R}^d$. Here, and for the remainder of this article, we use subscript i to denote the setting, and the subscript t to denote the instances observed within. Every time the setting is updated, $i \leftarrow i + 1$, and the observation counter t persists (e.g., if there were five instances in setting \mathcal{S}_1 , the first observation in the next setting \mathcal{S}_2 will be \mathbf{x}_6).

For every instance, the learner must decide whether or not to “query” *nature* for the true label y_t of the current instance \mathbf{x}_t , where $y_t \in \{\pm 1\}$ (for binary classification²). By *query* we mean the learner takes an action (*interact with nature*) and observes its outcome. Hence, we can think of \mathbf{x}_t as the contextual information associated with each action a_t . An observed feature vector, once queried, becomes a training instance to improve the learner. The training will be described in Section 2.5.

Let $Q_t \in \{0, 1\}$ denote the query indicator at time t . If a query is issued, i.e., $Q_t = 1$, the setting is updated ($i \leftarrow i + 1$), and the learner observes the label of the *queried instance*. It then updates

Algorithm 1: $a_{i+1} = \text{explorationPlanner}(\mathcal{S}_i, \mathcal{M}_i)$

```

1  $a_t \leftarrow \emptyset$  //initially idle
2 while  $t_{obs} > 0$  do
3   | observe  $\mathbf{x}_t$ 
4   |  $Q_t \leftarrow \text{isQuery}(\mathcal{M}_i, \mathbf{x}_t)$ 
5   | if  $Q_t = 1$  then
6   |   |  $a_t \leftarrow \text{Query}(\mathbf{x}_t)$ 
7   |   | break //continued at line 10
8   | end
9 end
10 if  $a_t = \emptyset$  then
11   |  $a_t \leftarrow \text{planning}(\mathcal{S}_i, \mathcal{M}_i)$ 
12 end
13 return  $a_{i+1} = a_t$ 
    //then execute  $a_{i+1}$  to get label

```

its hypothesis, taking into account the queried example (\mathbf{x}_i, y_i) as well as the previous hypothesis, which was learned over previous queries. Otherwise, i.e., $Q_t = 0$, the learner skips the current instance \mathbf{x}_t (meaning its label is not revealed) and continues to observe new instances from the current setting ($i \leftarrow i$).

Clearly, this constitutes a sequential decision process, which generates training examples for the learner. Since each interaction can require the learner to spend time and effort, i.e., labels are expensive to get, it is reasonable to set the objective of the decision process to be such that the learner *learns as much and as fast as it can*.

As a concrete example of this framework, consider our blocks-world environment. Here, a setting is a configuration of all the blocks on the table, while an instance \mathbf{x}_t is a feature vector encoding a possible placement location. The fovea sequentially provides possible placement locations, and, for each one, a new instance \mathbf{x}_t is observed. For each new instance in turn, the agent predicts the *outcome* of placement. Here, the binary outcome label indicates the success or failure of stacking. The label $y_t = 1$ indicates a stable placement, while the label $y_t = -1$ indicates an unstable placement.

After the action is taken, “nature” reveals a new setting \mathcal{S}_{i+1} and the agent *obtains*, through observation, the outcome and therefore the label, which will be used to improve its world model. In implementation, the agent obtains the outcome label by comparing two images of the configurations before and after the placement. This is possibly noisy, but usually correct.

2.4. PLANNING IN EXPLORATION

Our system has a set of adaptive classifiers to predict the block placement outcomes, which, together, constitute the world model \mathcal{M} . These obtain *knowledge* about the world, and a curiosity-drive causes the agent to desire to accumulate such knowledge (learning progress) as quickly as possible.

The agent is greedy in its pursuit of knowledge. For every instance \mathbf{x}_t observed during setting i , a *query condition* $Q_t \in \{0, 1\}$ is generated. The query condition is used to decide if this instance is worth querying for its label (outcome), based on the current model $\mathcal{M}_t = \mathcal{M}_i$. As soon as it encounters a true query condition, it executes the query, observes the outcome, and updates the

¹For more information on subspace planning, see related work in relational RL by Lang and Toussaint (2009).

²A more general framework would consider the multiclass and regression cases, which we leave for future extension.

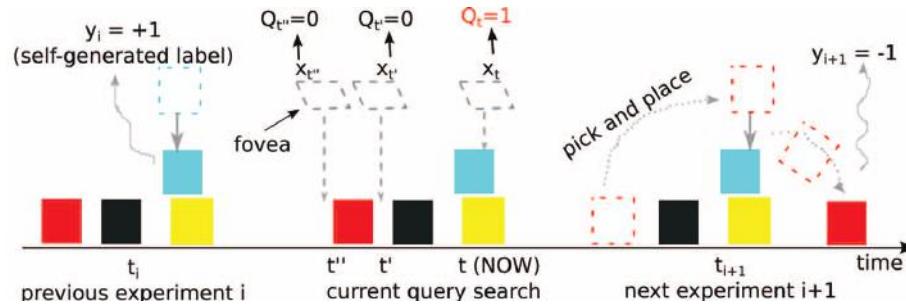


FIGURE 3 | A single robot-environment interaction, illustrating a setting change. Each pick and place “experiment” causes a change in setting. The outcome of the previous experiment was that the robot placed the blue block on top of the yellow block, and observed the label $+1$, corresponding to “stable.” Now (middle), the robot examines three fovea locations (t , t' and

t''), each of which involves a query. The query is false for t' and t'' , but true for t , and the robot immediately (greedily) grasps the furthest block, which happens to be the red one, and places it at the queried location. The action causes a change in setting to $i+1$ and the outcome -1 is observed (“unstable”).

model to \mathcal{M}_{i+1} . **Figure 3** illustrates this exploration behavior in our blocks-world environment.

But in the case where no instances in the setting are deemed to be valuable to query, the agent has to *plan*. In that case, the curiosity drive wants to quickly reach a new setting from which an instance worth querying *can* be observed. To decide which instances are worth querying, the agent simulates future experience of performing different actions from the current setting, and sees, for the simulated new settings, if the query condition becomes true at any point. If so, an intrinsic reward is placed at that transition. A true query condition in simulated experience becomes a binary curiosity reward indicating if an instance is worth exploring. By planning on the *induced* MDP with “phantom” reward function, the agent generates an efficient exploration policy whenever it needs to. These policies for reaching self-generated goals are the skills learned by the agent. Note that this curiosity reward is *instantaneous*, taking into account the current state of the learners, and not a previous learner. See **Algorithm 1** for a sketch of this process.

The planner can be implemented using any relevant MDP planning algorithms (Mausam and Kolobov, 2012), for instance, local methods (i.e., for the current state only) like UCT (Kocsis and Szepesvári, 2006), or global methods (for every state) like LSPI (Lagoudakis and Parr, 2003). In our implementation we use approximate policy iteration (LSPI, specifically the algorithm LSTDQ-Model), a global method, to allow the agent to choose between different states/heights (if several stacks are available) in each setting.

In the MDP constructed for our Katana blocks-world environment, the transition probabilities are derived from the adaptive classifiers. At planning time, we update the transition matrix $P(s'|s, a)$ for all state-action-state triplets as follows: $P(s'|s, a) = 0$ if $s' > s + 1$; $P(s'|s, a) = (1 + \hat{\Delta})/2$ if $s' = s + 1$; and $P(s'|s, a) = (1 - \hat{\Delta})/2/s$ if $s' \leq s$, with the prediction margin $\hat{\Delta}$ computed as the inner product between the contextual feature \mathbf{x} representing action a , and the linear weight vector \mathbf{w} of the predictor, i.e., $\hat{\Delta} = \mathbf{w} \cdot \mathbf{x}$ (more details will be provided in the next section). In other words, the transition probability to current height plus one is equal to the probability of a stable placement. It is zero for any height which is two or higher above the current one, and is

a uniform fraction of the probability of instability for the lower heights. Note that this is just an approximation, but it is good enough for effective planning to reach higher heights.

The next two sections describe our adaptive learners and the derivation of query condition, based on these learning models.

2.5. ONLINE LEARNERS

We focus on adaptive binary linear classifiers. There are multiple such classifiers in our system—one per height—but the discourse in this subsection will be with respect to a single classifier, for simplicity. For such a classifier, with weight vector $\mathbf{w}_t \in \mathbb{R}^d$, a classification of instance \mathbf{x}_t is then of the form $\hat{y}_t = \text{sign}(\mathbf{w}_t \cdot \mathbf{x}_t)$. The term $\hat{\Delta}_t = \mathbf{w}_t \cdot \mathbf{x}_t$ is often referred to as the *prediction margin* attained on instance \mathbf{x}_t , and the magnitude of the margin $|\hat{\Delta}_t|$ is a measure of confidence of the classifier in label prediction³.

In the setting of a developmental robot interacting with nature, *training* instances are generated in a biased manner. They are not independent and identically distributed—the sampling/query process depends on the learner’s adaptive model \mathcal{M}_t . However, their corresponding labels can be assumed to be generated from a linear stochastic model. Specifically, we make the following assumptions: 1) The labels $y_t \in \{-1, +1\}$ are realizations of independent random variables Y_t sampled from a stochastic source with a probability density function $P(Y_t|\mathbf{x}_t)$ continuous at all \mathbf{x}_t . This entails that, if $\Delta_t = \mathbb{E}[Y_t|\mathbf{x}_t] \in [-1, 1]$, then $\text{sign}(\Delta_t)$ is the Bayes optimal classification. 2) There exists a fixed but unknown vector $\mathbf{u} \in \mathbb{R}^d$ for which $\mathbf{u} \cdot \mathbf{x}_t = \Delta_t$ for all t . Hence \mathbf{u} is the Bayes optimal classifier under this noise model.

Note that when running our algorithms in a reproducing kernel Hilbert space (RKHS) \mathcal{H} with a universal kernel (Steinwart, 2002), the classifiers are implicitly non-linear, and Δ_t is well approximated by $f(\mathbf{x}_t)$, for some non-linear function $f \in \mathcal{H}$, hence assumption 2 becomes quite general.

The key elements in designing an online learning algorithm include the comparator class $\mathcal{U} \subseteq \mathbb{R}^d$, the loss function ℓ , and the update rule. For an arbitrary classifier $\mathbf{v} \in \mathcal{U}$, denote by $\ell(\mathbf{v}; \mathbf{x}_t, y_t)$ its non-negative *instantaneous* loss suffered on the

³Note that the terms *weight vector*, *linear hypothesis*, *classifier*, and *learner* are fairly interchangeable for the purposes of this article.

current example (\mathbf{x}_t, y_t) , and abbreviated by $\ell_t(\mathbf{v})$, i.e., $\ell_t(\mathbf{v}) = \ell(\mathbf{v}; \mathbf{x}_t, y_t)$. We define the total loss of an *adaptive* learner \mathcal{L} on a particular sequence of examples $\mathcal{D} = \{(\mathbf{x}_t, y_t)\}_{t=1}^T$ as $L_T(\mathcal{L}, \mathcal{D}) = \sum_{t=1}^T \ell_t(\mathbf{w}_t)$, and we also define the total loss of some (fixed) classifier \mathbf{v} as $L_T(\mathbf{v}, \mathcal{D}) = \sum_{t=1}^T \ell_t(\mathbf{v})$. A good learner that makes few online prediction mistakes also has small *relative* loss compared to the best linear hypothesis \mathbf{u} :

$$L_T(\mathcal{L}, \mathcal{D}) - \inf_{\mathbf{v} \in \mathcal{U}} L_T(\mathbf{v}, \mathcal{D}), \quad (1)$$

for any sequence \mathcal{D} . Since the online learner only observes one example at a time, the relative loss is the price of hiding future examples from the learner (Azoury and Warmuth, 2001). A desired analysis step in designing online learners is then to prove upper bounds on such a relative loss. This bound should grow sublinearly in T , so that it vanishes when T approaches infinity.

We use a modified version of the widely used regularized least square (RLS) classifier (Azoury and Warmuth, 2001; Cesa-Bianchi et al., 2009; Dekel et al., 2010)—a variant of the online ridge-regression algorithm—as our online learner. As the name suggests, this class of algorithms uses the squared loss function, and possesses a proven relative loss bound under our label noise model (Vovk, 2001; Dekel et al., 2012), with the desired sublinear growth. Established results for the algorithm will be used to derive our query condition (Section 2.6).

Given the sequence of queried (i.e., training) examples up to setting i , $\{(\mathbf{x}_j, y_j)\}_{j=1}^i$, the RLS classifier maintains a data correlation matrix, $A_i = I + \sum_{j=1}^{i-1} \mathbf{x}_j \mathbf{x}_j^\top$, with I the $d \times d$ identity matrix and $A_1 = I$. For the i -th queried instance \mathbf{x}_i , the weight vector can be updated as $\mathbf{w}_{i+1} = A_{i+1}^{-1}(A_i \mathbf{w}_i + y_i \mathbf{x}_i)$.

The inverse matrix A_{i+1}^{-1} can be updated incrementally using the Sherman-Morrison method,

$$A_{i+1}^{-1} = A_i^{-1} - \frac{\mathbf{b}_i \mathbf{b}_i^\top}{1 + c_i},$$

where

$$\mathbf{b}_i = A_i^{-1} \mathbf{x}_i$$

and

$$c_i = \mathbf{x}_i^\top A_i^{-1} \mathbf{x}_i = \mathbf{x}_i \cdot \mathbf{b}_i.$$

Using the fact that $A_{i+1}^{-1} \mathbf{x}_i = \mathbf{b}_i / (1 + c_i)$, the weight vector update is simplified as:

$$\mathbf{w}_{i+1} = \mathbf{w}_i + \frac{(y_i - \mathbf{w}_i \cdot \mathbf{x}_i)}{1 + c_i} \mathbf{b}_i.$$

An implementation-efficient pseudocode of this modified RLS update rule is presented in **Algorithm 2**.

Algorithm 2: $\mathcal{M}_{i+1} = \text{modifiedRLS}(i, \mathbf{x}_i, y_i, A_i^{-1}, \mathbf{w}_i)$

```

1 if  $i = 0$  then
2    $A_1^{-1} = I // d \times d$  matrix
3    $\mathbf{w}_1 = \mathbf{0} // d \times 1$  vector
4 else
5   //  $\mathbf{b}_i, c_i$ : useful intermediate terms
6    $\mathbf{b}_i = A_i^{-1} \mathbf{x}_i; c_i = \mathbf{x}_i \cdot \mathbf{b}_i;$ 
7   // Projection step if  $\hat{\Delta}_i > 1$ 
8    $\hat{\Delta}_i = \mathbf{w}_i \cdot \mathbf{x}_i;$ 
9    $\bar{\mathbf{w}}_i = \mathbf{w}_i - \text{sign}(\hat{\Delta}_i) \max\{0, \frac{\hat{\Delta}_i - 1}{c_i}\} \mathbf{b}_i;$ 
10  // Sherman-Morrison incremental update
11   $A_{i+1}^{-1} = A_i^{-1} - \frac{\mathbf{b}_i \mathbf{b}_i^\top}{1 + c_i}; \mathbf{w}_{i+1} = \bar{\mathbf{w}}_i + \frac{(y_i - \bar{\mathbf{w}}_i \cdot \mathbf{x}_i)}{1 + c_i} \mathbf{b}_i;$ 
12 end
13 return  $\mathcal{M}_{i+1} = (A_{i+1}^{-1}, \mathbf{w}_{i+1})$ 
//  $\mathcal{M}_{i+1}$ : updated model after experiment  $i$ 

```

2.6. QUERY CONDITION

Our query condition is greatly inspired by work in selective sampling, a “stream-based” setting of active learning (Atlas et al., 1989; Freund et al., 1997). In selective sampling, the learner has access to an incremental stream of inputs and has to choose, for each datum in order, whether to query its label or not. State of the art methods in selective sampling, with theoretical performance guarantees, include BBQ (Orabona and Cesa-Bianchi, 2011) and DGS (Dekel et al., 2012). These methods also use variants of the RLS algorithm (Azoury and Warmuth, 2001; Vovk, 2001; Auer, 2003; Cesa-Bianchi et al., 2005; Cesa-Bianchi and Lugosi, 2006; Cavallanti et al., 2008; Strehl and Littman, 2008; Cesa-Bianchi et al., 2009), and maintain a data correlation matrix to calculate a confidence interval or uncertainty level in their prediction, which is essentially an estimate of the variance of the RLS margin for the current instance.

The query condition must indicate when the outcome is statistically known or unknown. Here we derive a query condition for this purpose, based on the expected learning progress. Essentially, when the learner is certain in what it predicts, it can ignore the instance, since, with high probability, its learning model will not get updated much on this example if it is queried. Inversely, only those instances that the learner is uncertain in its prediction are worth querying for labels, since the model of the learner will undergo a large update on such training examples.

The following lemma from Orabona and Cesa-Bianchi (2011) defines χ_t , the *uncertainty level*, or *confidence interval* of the RLS prediction.

Lemma 1. Let $\delta \in (0, 1]$ be a confidence level parameter, $h_{\delta, \mathbf{u}}(t)$ be a function of the form

$$h_{\delta, \mathbf{u}}(t) = \|\mathbf{u}\|^2 + 4 \sum_{k=1}^i r_k + 36 \log \frac{t}{\delta},$$

where $\|\mathbf{u}\|$ is the unknown squared norm of the optimal Bayes classifier, and $r_i = \mathbf{x}_i^\top A_{i+1}^{-1} \mathbf{x}_i$.

Now, define $\chi_t = \sqrt{c_t h_{\delta, \mathbf{u}}(t)}$ with $c_t = \mathbf{x}_t^\top A_{i+1}^{-1} \mathbf{x}_t$. With probability at least $1 - \delta$, the following inequality holds simultaneously for all t :

$$|\Delta_t - \widehat{\Delta}_t| \leq \chi_t.$$

This inequality can be rewritten as,

$$\Delta_t \widehat{\Delta}_t \geq \frac{\Delta_t^2 + \widehat{\Delta}_t^2 - \chi_t^2}{2} \geq \frac{\widehat{\Delta}_t^2 - \chi_t^2}{2},$$

which essentially implies that if $|\widehat{\Delta}_t| > \chi_t$, the learner is **certain** (with probability at least $1 - \delta$) that $\widehat{\Delta}_t$ and Δ_t have the same sign (i.e., $\widehat{\Delta}_t \Delta_t > 0$), and **there is no need to query for the true label**. Inversely, when $|\widehat{\Delta}_t| \leq \chi_t$, the learner is **uncertain** about its prediction, and **it needs to issue a query**. Formally, the query condition is stated as follows:

$$\text{isQuery}(\mathcal{M}_{i+1}, \mathbf{x}_t) : Q_t \leftarrow [\chi_t > |\widehat{\Delta}_t|],$$

where $[\cdot]$ denotes the indicator function of the enclosed event.

Now, from Lemma 1 we also have $|\Delta_t| \leq |\widehat{\Delta}_t| + \chi_t$. Combined with the query condition derived above, we have $|\Delta_t| \leq 2\chi_t$ with probability at least $1 - \delta$ when a query is issued. When the magnitude $|\Delta_t|$ of the optimal prediction margin is small, the instance label is almost certainly noise, i.e., the prediction is nearly a random guess. These instances are “hard” or even “impossible” to learn, and the learner should instead focus on other instances that it can improve its prediction capability. We derive another query condition to reflect this insight, by enforcing another threshold θ on the uncertainty level,

$$\text{isQuery}(\mathcal{M}_{i+1}, \mathbf{x}_t) : Q_t \leftarrow [[\chi_t > |\widehat{\Delta}_t|] \wedge [\chi_t > \theta]]. \quad (2)$$

In implementation, a surrogate or proxy function is used to avoid dependency on the optimal yet unknown \mathbf{u} . This takes the form,

$$\chi_t = \alpha \sqrt{c_t h(t)},$$

where α is a tunable positive parameter, and

$$h(t) = \log(1 + i)$$

is a simplification of $h_{\delta, \mathbf{u}}(t)$. Importantly, the confidence interval does not depend on the squared norm of the optimal but unknown Bayes classifier \mathbf{u} . See Dekel et al. (2012) Equation (12) and Lemma 7, notice the additional assumption of $\|\mathbf{u}\| \leq 1$. See also Orabona and Cesa-Bianchi (2011) Algorithm 2 for another proxy function.

3. RESULTS

In all implementations we used the following parameter values: discount factor $\gamma = 0.95$, and query condition scaling

factor $\alpha = 1$. The confidence-interval threshold $\theta = 0.01$ for simulations, while $\theta = 0.1$ was used in the real robot experiments.

3.1. SIMULATED BLOCKS-WORLD ENVIRONMENT

We designed a stripped-down simulated version of the actual blocks-world, in order to test our system. In simulation, thousands of trials can be run, which would take far too long on the real robot. Of course we cannot capture all aspects of the real-world robot setting, but we can capture enough so that the insights and conclusions arising from simulated results suffice to evaluate our system’s performance.

The simulated environment also allows us to use any number of blocks and any number of features. For any configuration of blocks, some set of heights will be available for the agent to place upon, corresponding to the heights of the top blocks in the stack(s), and height zero. In the simulation, we use eight blocks, and 21 features. Each height’s feature vector is of length 21 bits, with only one bit set. All 21 feature vectors are available for each available height. The agent must select one of them. Unlike the actual robot setting, in simulation, the features do not correspond to any physical aspect of the simulated world. In simulation, each of the 21 features are associated with a different probability of stability, which is randomly generated.

Each possible height s has a *different* weight vector \mathbf{u}^s , which is the randomly generated “true model” for the result of placing a block upon it. This was done in order to generate simulated block placement outcomes in an easy-to-implement way. There are 21 components⁴ of each \mathbf{u}^s , which are randomly generated in the range $[-1, 1]$. An outcome (stable/falling) is generated using the corresponding height’s true (probabilistic) model, where the actual outcome label $\text{sign}(\mathbf{u} \cdot \mathbf{x}_t)$ is flipped with probability $\frac{1 - |\mathbf{u} \cdot \mathbf{x}_t|}{2}$. For the purpose of generating orderly plots in Section 3.2, we re-order the 21 feature vectors of each height in ascending order of their likelihood of stability, then re-assign their feature indices from 1 to 21. Thus, the smaller the feature index, the lower likelihood the placement will be stable. For an outcome of falling, there is a chance that the entire stack underneath the placement position collapses, in which case all blocks in that stack are reset to height one.

The eight blocks’ configuration is represented by vector \mathbf{q} . The absolute value of each element $|q_j|$ is the height of the corresponding block j . We set $\text{sign}(q_j) = -1$ if block j is occluded (stacked upon), while $\text{sign}(q_j) = 1$ means block j is on the top of its stack, which means its both graspable and another block can be placed upon it. The set of different positive elements of \mathbf{q} constitute the set of current available states (heights to place upon) in addition to height zero (which is always available). For example, vector $\mathbf{q} = (-1, -2, -3, 4, -1, 4, -2, -3)$ means the configuration has two different stacks of height four, having block IDs 4 and 6 on top of the two stacks. Here, the set of available placement heights is height zero and height four.

⁴To allow generalization in learning, each weight vector is extended with one extra bias component, corresponding to an extra augmented feature of 1. Thus, $|\mathbf{u}| = |\mathbf{w}| = |\mathbf{x}_t| = 22$ in the implementation of the simulated environment, and $|\mathbf{w}| = |\mathbf{x}_t| = 7$ in the implementation on the real robot.

After selecting the state and action, the agent picks an “available” top block, and “places it.” By available, we mean it is the top block of another stack. Another block in the stack (if any) of the block that is grasped becomes a top block. If the placement is stable, the highest block in the placement stack has its sign reversed, and the placed block becomes the top block of that stack. If the placement outcome turns out to be unstable, a “toppling” event occurs, where one randomly selected block in the stack of placement, with a lower height, becomes a top block of the remaining stack, with blocks below unchanged. The (unsuccessfully) placed block and the other, higher blocks in the stack topple to the surface, and their values are all set to +1.

3.2. RESULTS IN SIMULATED BLOCKS-WORLD

Figure 4 shows the *averaged* exploration behavior of our system over time, for all different heights. “Direct exploration” refers to settings where the query condition is true, while “planning experience” refers to settings where the algorithm has to execute a planned action (since the query condition is always false for that setting). On the *y*-axis, “cumulative experience” is a count of the number of times these types of actions are generated. The different colored lines indicate different heights. The vertical lines are from a single run, and indicate when, during that run, the learner switches from direct exploration of one height to planning exploration of higher heights.

These plots show the developmental stages of the learning agent, where easier problems, such as direct exploration at height one, are learned first, and more difficult problems are learned later. They also show cumulative learning, as the acquired knowledge at lower heights is exploited for planning, and this planning helps the agent get to the higher heights, in order to acquire more knowledge. The difficulty of this problem is shown by the time the learner needs to spend to fully explore its environment, especially in achieving the highest heights. For instance, to even get to height six to do experiments, the agent first needs to stack

blocks from lower heights each time the stack collapses, which is a regular occurrence.

The agent does not necessarily explore a single height until everything at that height is statistically known. There are sometimes situations where several heights worth exploring are available simultaneously in the environment. In such cases, the agent starts with the height having the largest “future exploration value” as estimated by LSPI. The planning step helps to trade off “easy-to-get” small learning progress rewards with “harder-to-get” larger ones. As shown in **Figure 4**, the exploration at higher heights does, in fact, start before the direct exploration of lower heights terminates.

Figure 5 shows the learning progress, measured with Kullback-Leibler (KL) divergence between the learned models and the true models. These distances tend to diminish exponentially with experience, and they diminish faster at lower heights, where experience is easier to get. When each line in the graph saturates, it corresponds to the associated knowledge being “known” and ready for exploitation in planning. The saturation levels are non-zero due to the noise level in the training labels, the query condition scaling factor α , and the confidence-interval threshold θ .

Figure 6 shows how the *exploration focus* changes over time, for height one. In each subgraph row, the figure on the left shows the distribution of the experience up until the timestep in the subfigure title. The shaded area between the two vertical lines represents the “unknown” region of input features that is deemed to still be worth exploring. This will be the “exploration focus” of the agent, in subsequent interactions. Regions outside of this shaded area are considered “known” by the learning agent, and not worth exploring any more. Going from the top to the bottom of **Figure 6**, note that the query region shrinks with the amount of experience. Additionally, note that the middle features, associated with the most uncertain outcomes (as mentioned in Section 3.1) stay interesting longer than the others.

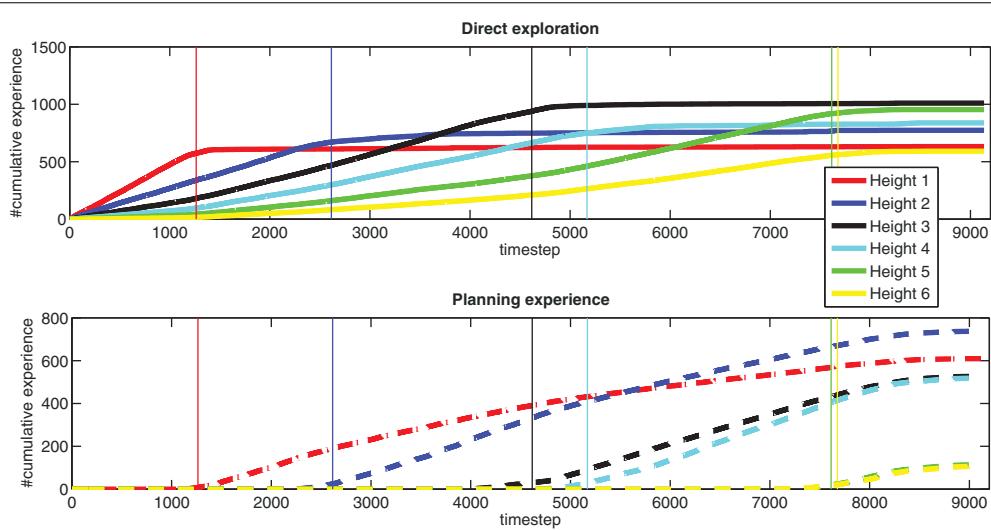


FIGURE 4 | Exploration history (averaged over 10 runs).

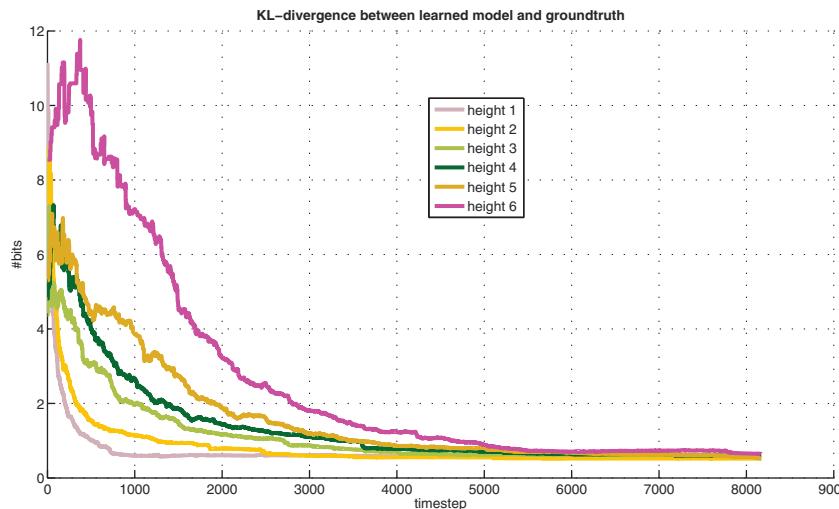


FIGURE 5 | KL-divergence between learned models and ground-truth models (averaged over 10 runs). Best viewed in color.

An interesting observation that is worth elaborating on, is as follows. At timestep #2000, when every prediction is statistically known, the agent starts to exploit the acquired knowledge for planning (i.e., taking its estimated “best” action #16 to reach height two). It also keeps on refining the learned model, which reveals, as a result of generalization in learning that the optimal action (i.e., the most stable placement position) is action #20 instead. Afterwards, the agent switches its optimal policy for this height, as shown in timesteps #3000 and #4000.

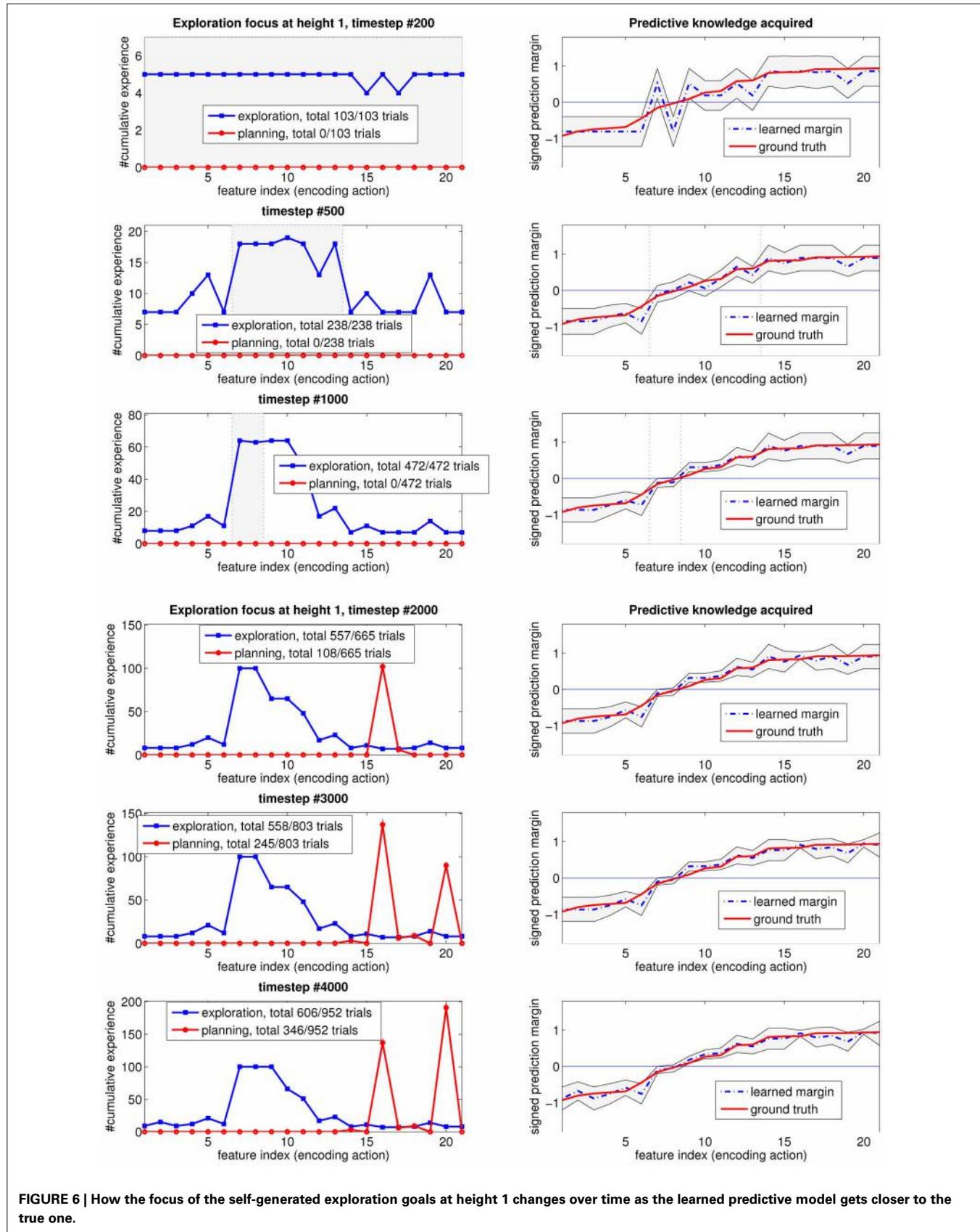
The plots on the right shows the learned predictive model (blue dashed lines), with two thin black lines representing the confidence intervals for each prediction. As more data are observed, the associated confidence interval will shrink, reflecting the learning progress. Note that as a result of generalization, the neighboring area of the input feature space also gets improved, indirectly, in its confidence interval. Recall that we re-arranged the input feature indices so that their prediction margin (hence, probability of stable/unstable outcomes) are in ascending order.

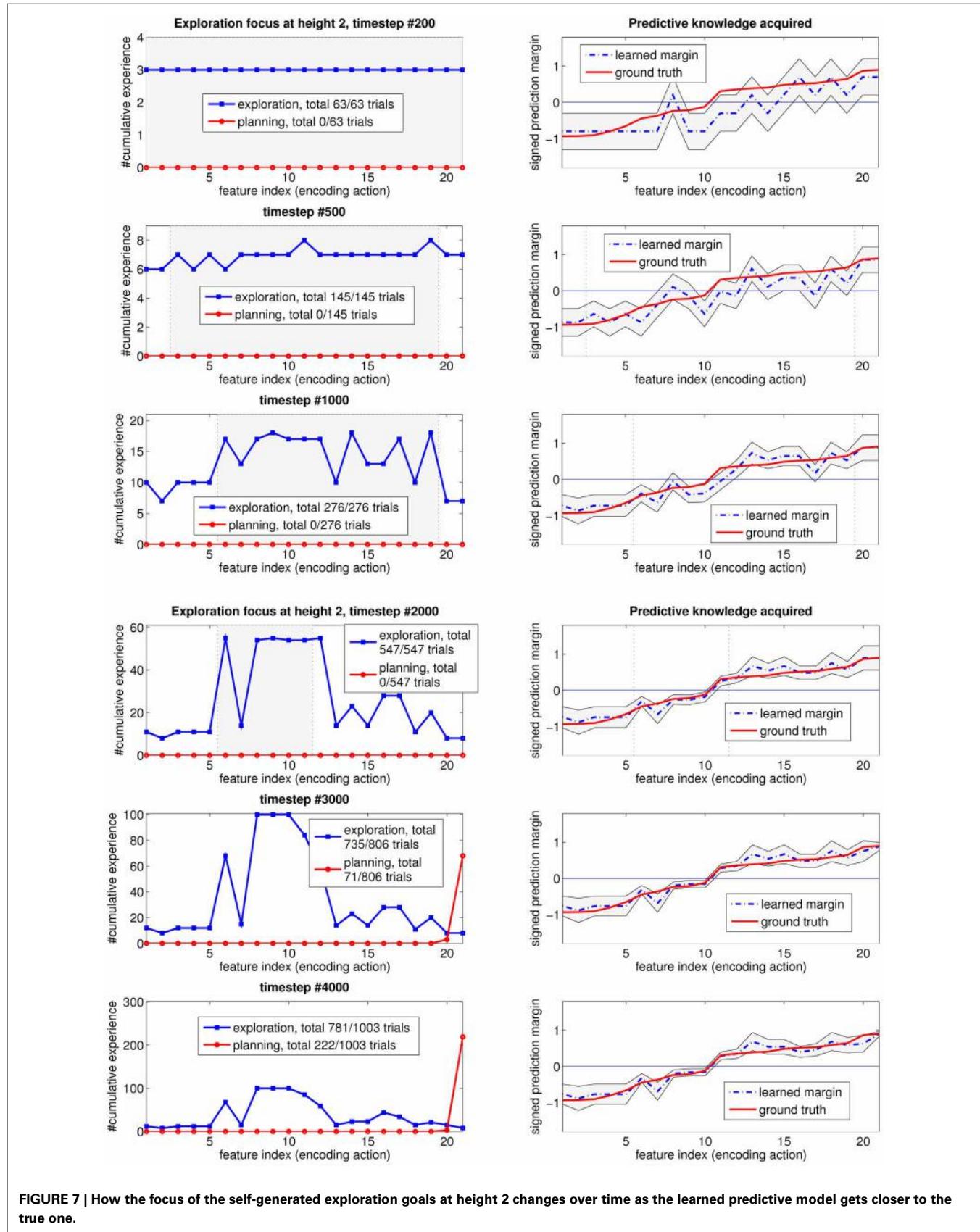
A measure of the difficulty of a learning problem is the sample complexity needed to achieve some desired level of confidence. The shaded regions (i.e., “unknown” and worth exploring) shrink with experience, toward the input feature values with small prediction margin ground-truth. These feature values correspond to the input subspace with prediction outcomes close to noise, i.e., hard to predict. However, these instances lying close to the decision boundary are the most informative instances for constructing a good decision plan. Our system first explores much of the input space, then quickly shifts its attention to this “hard-to-learn” input region, where most of its exploration effort is spent. As a result, the learned predictive model gets closer to the true model over time. Note that for “known” regions outside the shaded area, even though the number of experiences is small, and the confidence interval (i.e., uncertainty level) is large, the learning algorithm is still confident that its prediction (sign of the margin) is close to the optimal one with high probability. Thus, these regions are not worth exploring any more.

The same exploration behavior is observed when we analyze the data for other heights, as shown in **Figure 7** for height two, and **Figure 8** for the first six heights when exploration terminates. In all the experiments, the agent first explores the whole input feature space, then focuses on subspaces of input features that are informative but for which high confidence is hard to achieve, then on features that are useful for planning. This typically occurs for each height in turn. As a result of learning how to plan, which necessarily entails reliably transitioning from one state (height) to another, the skill of block stacking is achieved.

To further analyze the effectiveness of our method, we compare its performance to three other methods. The comparison measure is the KL-divergence with respect to the true model. The first method simply is uniform random action selection, which results in undirected, babbling-like, behavior. The second method, which we call Conf (Ngo et al., 2012), uses confidence intervals χ_t of the prediction margin directly as phantom rewards to generate the exploration policy through planning. Intuitively, this is also an informed exploration method since it promotes exploration in parts of the environment with high uncertainty. The main difference is the confidence intervals are used themselves as rewards, instead of using a query condition. The third method is a variant of our proposed method, but the exploration policy is updated (i.e., planning) after every 10 observations, instead of on-demand whenever exploration planning is invoked. We denote this variant as Q10, and our proposed method as Q1.

The results are shown in **Figure 9**, with each subgraph showing the KL-divergence between learned models and their ground-truth at each timestep. Inspecting carefully the subgraph for height one and two, we see that Q1 gets close to the true model exponentially fast in the first 1000 timesteps, then saturates. The random method, on the other hand, though making much slower progress than Q1 and Q10 in the first 1000 timesteps, keeps improving its learned models and achieves the best models for height one and two, among the four methods. However, for the other five higher heights, its learned models are much worse





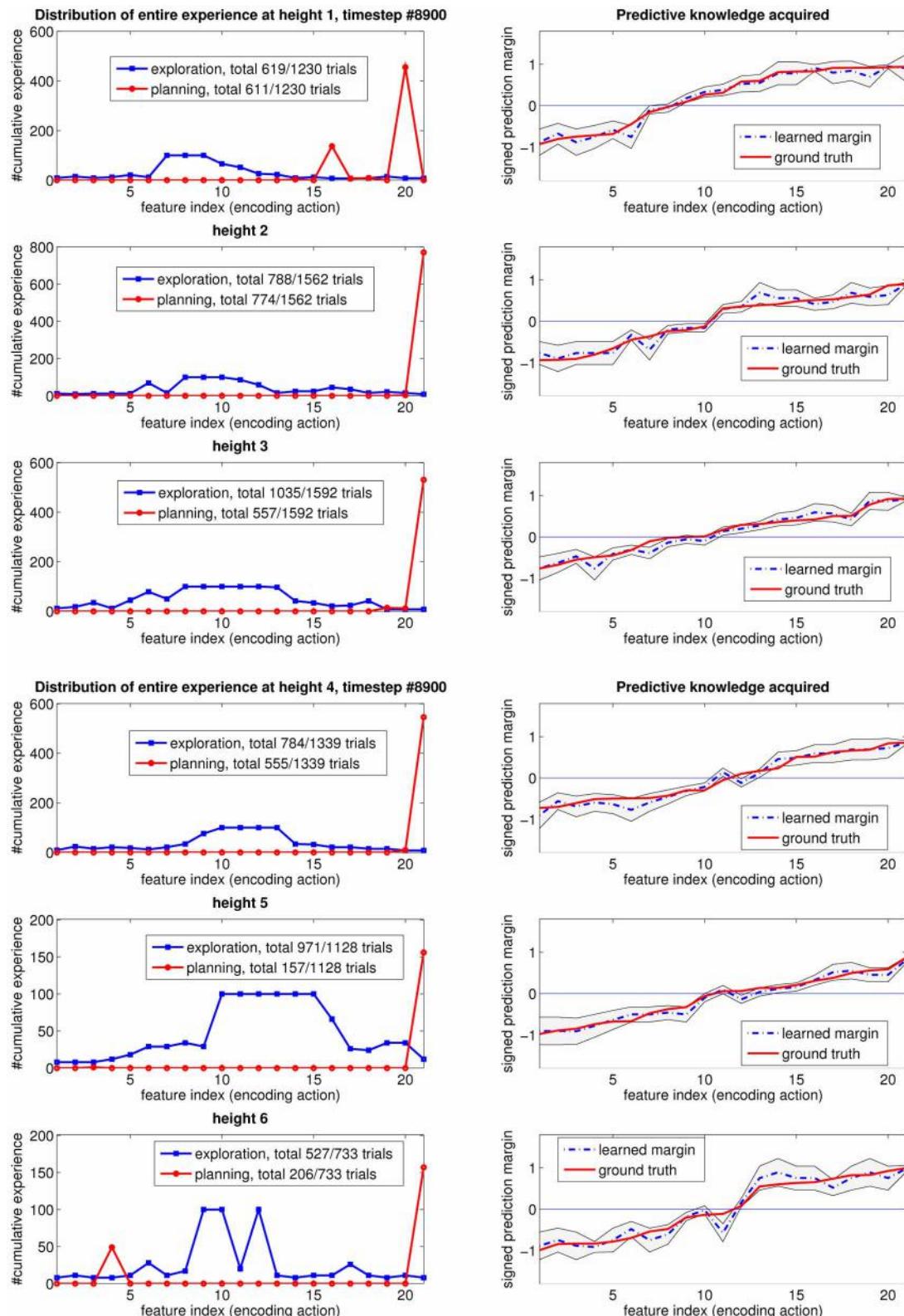


FIGURE 8 | Experience distribution after the last timestep (learning has completed) for heights 1–6.

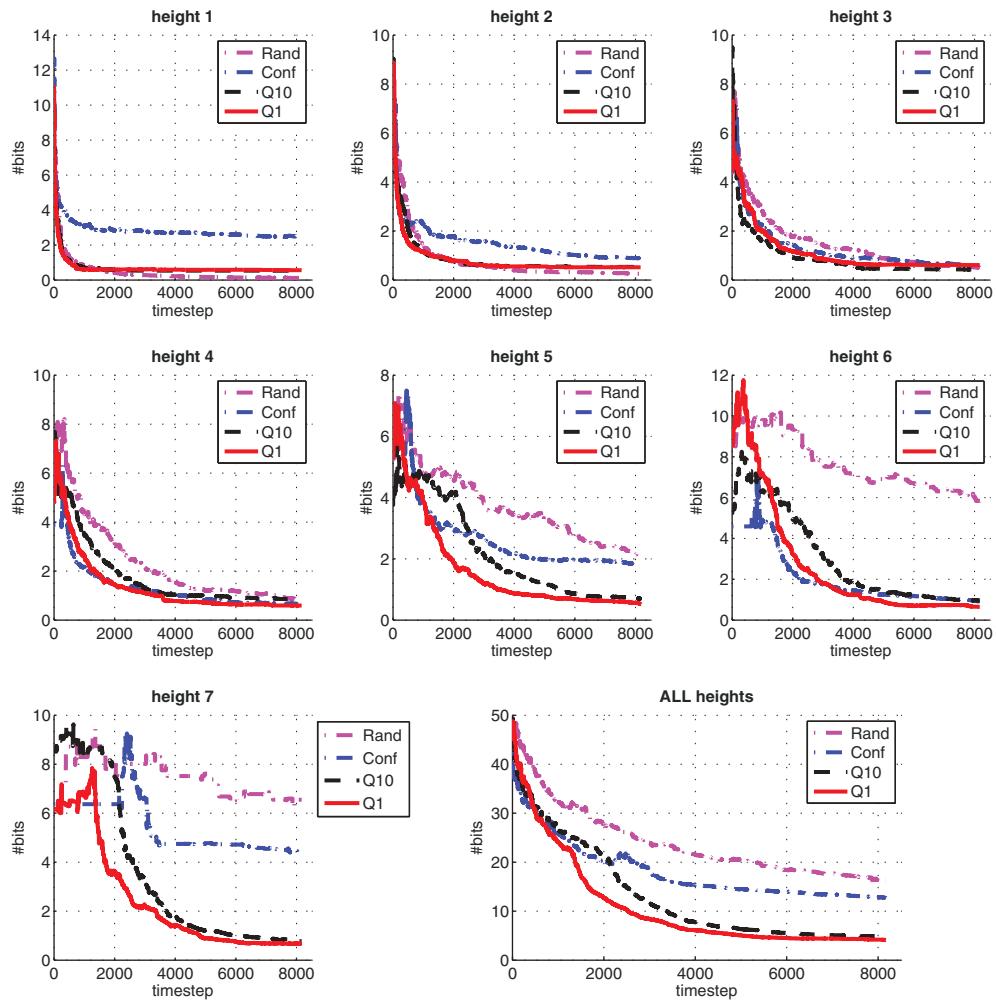


FIGURE 9 | A comparison of exploration methods in terms of the KL-divergence between the learned predictive models at each time step and their ground-truth models. Results are averaged over 10 runs.

compared to the rest. This can be explained by the fact that the blocks-world environment naturally generates unbalanced experience distribution among all the states under random action selection, and lower heights will get much more learning experience compared to higher ones. This undirected exploration behavior makes random exploration the least efficient method compared to the other three (informed) exploration methods, as shown in the overall results in the last subgraph at the bottom-right corner. The confidence-based method performs much better than random method, but is still inferior compared to query-based methods Q1 and Q10. The overall performance of Q1 is the best, closely followed by Q10, which is less efficient due to less frequent planning updates.

3.3. RESULTS ON THE REAL ROBOT

Now, we show the learning behavior on the real robot. Figures 10–12 show a snippet of experience consisting of 12 consecutive experiment sequences. In each frame, one should focus on the configuration of the blocks in the workspace and track

the changes from the previous frame. Each sequence starts with i) a fovea-based search for the desired placement in the current block configuration (i.e., either the query condition returns “unknown” or the best planned action is selected), as shown in the first column, followed by ii) an action picking a block unrelated to the placement experiment (second column), then iii) placing the block at the desired height, orientation, and relative position with respect to the stack below (third column). The sequence ends with an observation process to self-generate the label (last column). The end of one sequence is also the beginning of the next sequence. Since the robot has already had some prior experience before continuing from sequence #1 of the snippet, it now focuses on exploring height two. Specifically, from all the 12 sequences, we find that the robot gradually shifted its attention (from the second sequence in Figure 11 to the second last sequence in Figure 12 to trying actions A_3 and A_4 (corresponding to relative placement positions with two and three bits set), which are actually the actions with the most uncertain placement outcomes among the six actions. Note that with tower height four,

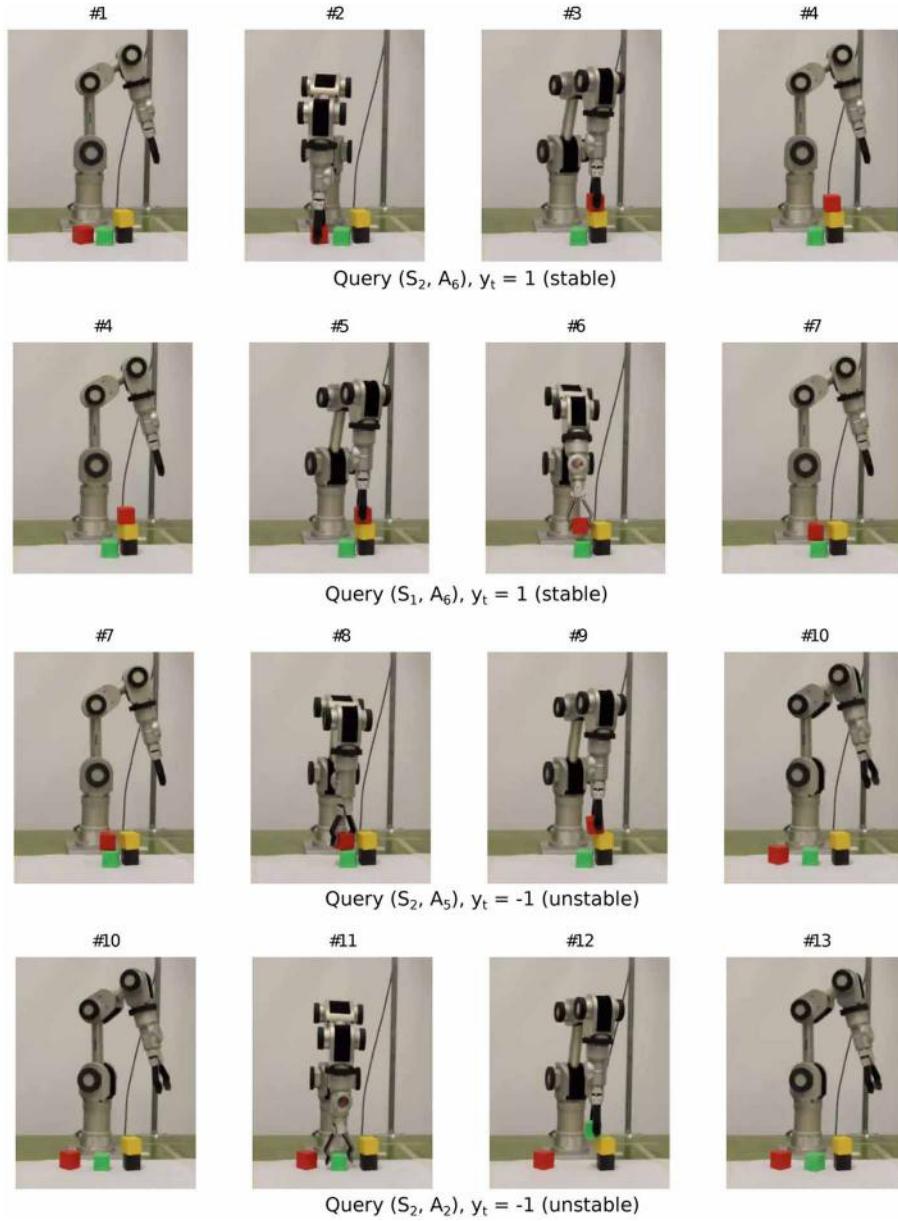


FIGURE 10 | Sample query sequence on real robot (1/3).

the robot arm does not have many feasible workspace points for the pick and place task. Hence we limit the maximum height to three.

Figure 13 shows the predictive models the Katana robot arm acquired in a single run with 30 interactions (see demo video at www.idsia.ch/~ngo/frontiers2013/katana_curious.html; the last 12 interactions shown in **Figures 10–12** start from 1:52).

Figure 14 shows a “tricky” situation for the robot, which it can overcome if it has learned the model well. Here, the robot must demonstrate its block stacking skill, as an externally imposed goal.

4. DISCUSSION

4.1. SYSTEMATIC EXPLORATION

This work was conceived with *pure exploration* in mind, which is contrasted with the treatment of exploration in classical RL. There, exploration is discussed in terms of the *exploration-exploitation tradeoff*. On the one hand, the agent should *exploit* the acquired knowledge by selecting the current best (greedy) action, thereby not spending too much time in low-value areas of the state space. On the other hand, it needs to *explore* promising actions to improve its estimation of the value function, or to build a more accurate model of the environment.

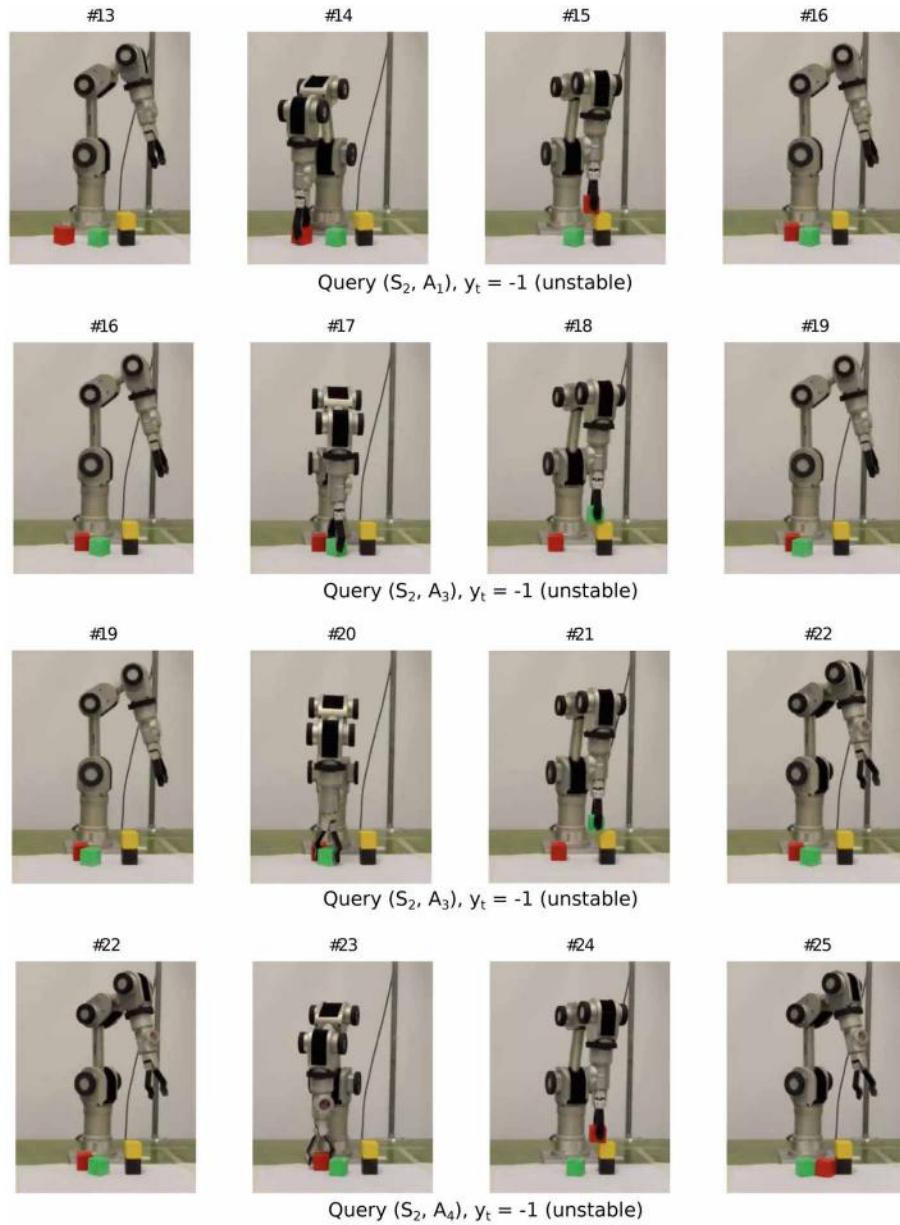


FIGURE 11 | Sample query sequence on real robot (2/3).

The most widely used method for balancing exploration and exploitation is the ϵ -greedy algorithm (Watkins and Dayan, 1992; Sutton and Barto, 1998). At each state, with probability of $1 - \epsilon$ the agent selects the greedy action with respect to the estimated value function, and with a small probability of ϵ it selects a *random* action for exploration. *Optimistic initialization* is another common method for exploration (Sutton and Barto, 1998). By initializing the value function for all states with high values, the agent will try to reach less visited states until their values converge to near-optimal ones, which is much lower than the initial values. The initial values strongly affect the exploration time. Progress-driven artificial curiosity is a more general method for balancing

exploration and exploitation which 1. removes the reliance on randomness—the exploration is *informed*, instead of relying on randomness (*uninformed*), and 2. promotes exploration of states where learning can occur over states where not much can be learned. To contrast, in optimistic initialization, every state is equally worth exploring.

Somewhat recently, several algorithms modifying optimistic initialization have been proposed that guarantee to find near-optimal external policies in a polynomial number of time steps (PAC-MDP). These algorithms, such as E^3 (Kearns and Singh, 2002) and R-max (Brafman and Tennenholz, 2003), maintain a counter for the number of times each state-action pair is

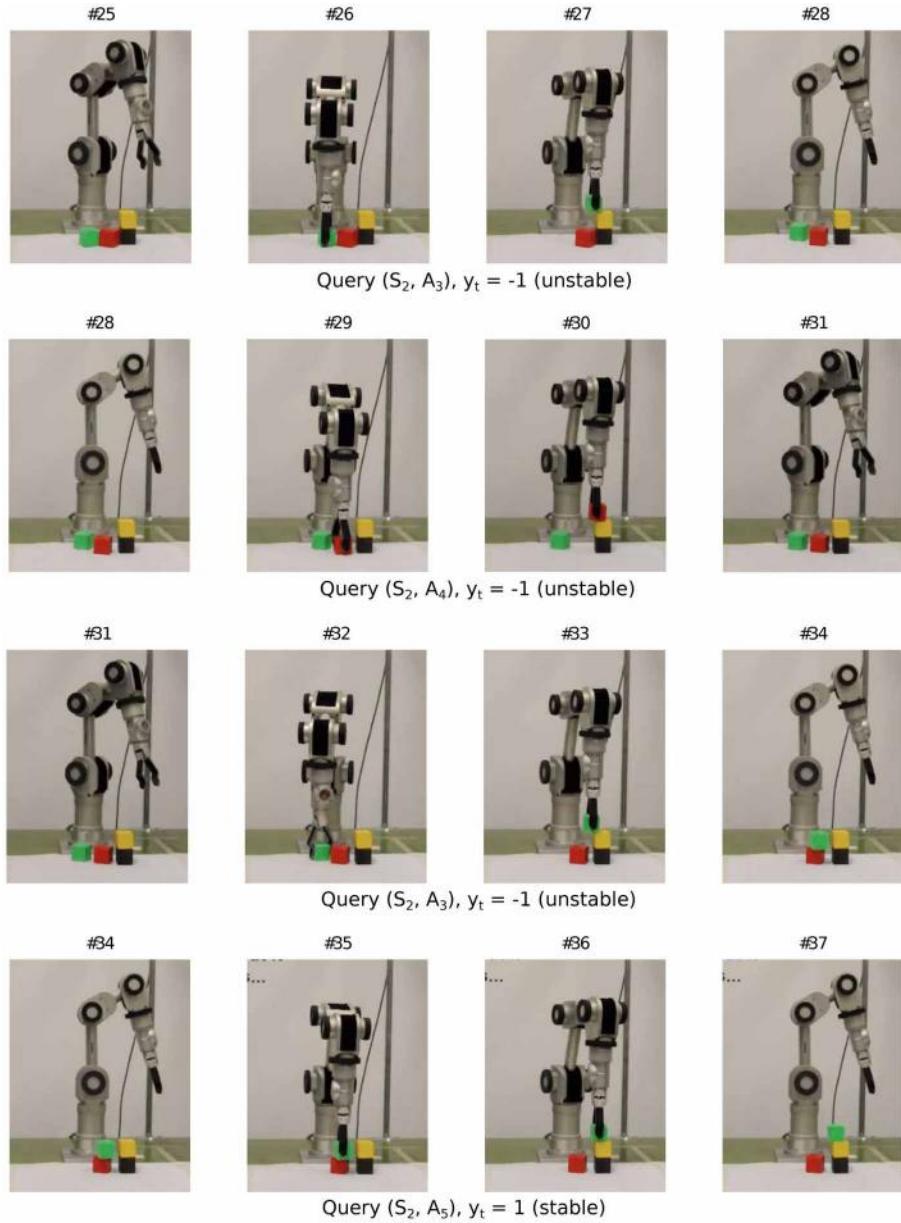


FIGURE 12 | Sample query sequence on real robot (3/3).

tried. When this number exceeds some threshold, the estimated state-action value is quite accurate, and the state-action pair will be considered “known”—thus with high probability the greedy action will be near-optimal (exploitation). Otherwise, the value is replaced with a highly optimistic one, encouraging the agent to explore such “less-selected” state-action pairs. Recent work in this model-based line of research extends R-max in several aspects. Rao and Whiteson (2012) give a better estimate of the optimistic reward using a weighted average between experienced and optimistic ones, resulting in the V-MAX algorithm that is capable of exploiting its experience more quickly. Lopes et al. (2012) propose to replace the counter of visits to a state with expected learning

progress based on leave-one-out cross-validation on the whole interaction history. Our method for estimating learning progress is, in contrast, instantaneous and online. Furthermore, it is able to generalize across different actions, instead of treating them separately.

The common theme in many intrinsically motivated RL approaches is that the estimated learning progress is used as secondary to external rewards. The purpose of the behavior (i.e., the policy) of the agent has a goal of achieving external rewards. Exceptions include, for instance, Şimşek and Barto (2006), where the agent’s behavior is based on a second value function using an intrinsic reward signal, which is

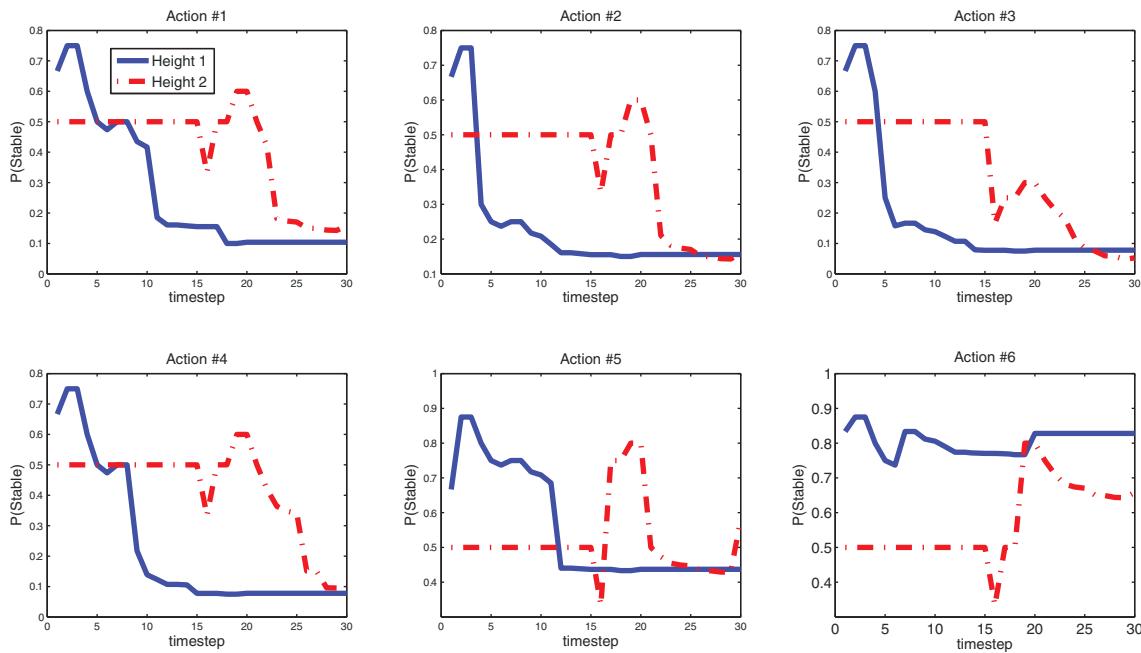


FIGURE 13 | Learning progress of the Katana robot arm’s predictive models at height 1 and 2 after 30 settings. Action 1 (no bits set) is the most unstable. Action 6 (all bits set) is the most stable. See earlier discussion on the features and **Figure 2**.



FIGURE 14 | A “tricky” situation to test the robot’s stacking skill. We show this case to illustrate the value of exploring to learn how the world works. Consider the robot is faced with a task to build a stack of blocks as fast as possible from this initial setting. Given its learned model of the world, the robot will decide to start stacking from height 1 instead of height 2, as with high probability the stack of two blocks will fall after placing another block upon them.

calculated based on the changes in value estimates of external rewards.

Besides our preceding work (Ngo et al., 2012), which this work is an extension of, some recent work in the pure exploration setting also uses planning. Yi et al. (2011) develop a

theoretically optimal framework based on the Bayesian methods, in which the agent aims to maximize the information gain in estimating the distribution of model parameters. An approximate, tractable solution based on Dynamic Programming is also described. Hester and Stone (2012) present results on simulated environments, where *two* progress-based intrinsic reward signals are used for exploration: one based on the variance in predictions of a decision tree model, and one based on the “novelty” of the state-action pair, to promote the exploration focus to shift toward more complex situations. In our system, we use a *single* curiosity reward signal based on the derived query condition, and our approach has been shown to be more effective than the previous variance-based approaches, since observations with large variance will not be worth querying *if* the learner is confident about its predictions.

In all the aforementioned work with pure exploration, planning is used to generate exploration policies, which must be invoked at every timestep. It has been observed (Gordon and Ahissar, 2011; Luciw et al., 2011) that quickly learning agents do not update their exploration policies fast enough to achieve the intrinsic rewards they expect to achieve. In such cases, learning progress-based exploration is no better than random action selection or various simple heuristics. In other words, the update speed of the policy generation must be much greater than the learning speed of the underlying learner. This can be computationally demanding. It can also be wasteful, when the intrinsic reward that the agent plans to achieve is, while non-zero, quite small.

Our approach allows the agent to choose the most informative observations (possibly several steps ahead) to sample, and *only* invoke expensive planning when the current situation is

already “known.” A statistically “known” prediction means the agent knows with high probability that its prediction is almost as correct as that of the Bayes optimal predictor. Due to this approach, the computational demands are reduced compared to a regular planner, and further, the agent will know when to stop its planning efforts—when everything is “known.”

4.2. CONCLUSION

Goal-driven exploration is very common in the traditional RL setting. In the pure-exploration setting, self-generated goals are needed. The agent described here generates goals based on its confidence in its predictions about how the environment reacts to its actions. When a state-action outcome is statistically unknown, the environment setting where that experience can be sampled becomes a goal. The agent uses planning to manipulate the environment so that the goal is quickly reached. Without planning, only local, myopic exploration behavior can be achieved. The result is a sample-efficient, curiosity-driven, exploration behavior, which exhibits developmental stages, continual learning, and skill acquisition, in an intrinsically-motivated playful agent. Key characteristics of our proposed framework include: a mechanism of informed exploration (with no randomness involved), a clear distinction between direct and planned exploration (i.e., planning is done only when all local instances are statistically known), and a mathematically-solid way of deciding when to stop learning something and when to seek out something new to learn.

ACKNOWLEDGMENTS

We would like to thank the reviewers for their very useful comments that helped improve this paper.

FUNDING

This work was funded through the 7th framework program of the EU under grants #231722 (IM-CLeVeR project) and #270247 (NeuralDynamics project).

REFERENCES

- Asada, M., Hosoda, K., Kuniyoshi, Y., Ishiguro, H., Inui, T., Yoshikawa, Y., et al. (2009). Cognitive developmental robotics: a survey. *IEEE Trans. Auton. Ment. Dev.* 1, 12–34. doi: 10.1109/TAMD.2009.2021702
- Atlas, L. E., Cohn, D. A., and Ladner, R. E. (1989). “Training connectionist networks with queries and selective sampling,” in *Advances in Neural Information Processing Systems 2*, ed D. S. Touretzky (Denver, CO: Morgan Kaufmann Publishers Inc.), 566–573.
- Auer, P. (2003). Using confidence bounds for exploitation-exploration trade-offs. *J. Mach. Learn. Res.* 3, 397–422. doi: 10.1162/153244303321897663
- Azoury, K. S., and Warmuth, M. K. (2001). Relative loss bounds for on-line density estimation with the exponential family of distributions. *J. Mach. Learn. Res.* 43, 211–246. doi: 10.1023/A:1010896012157
- Barto, A., Singh, S., and Chentanez, N. (2004). “Intrinsically motivated learning of hierarchical collections of skills,” in *Proceedings of International Conference on Development and Learning (ICDL)*, (San Diego, CA), 112–119.
- Bellman, R. (1957). A markovian decision process. *J. Math. Mech.* 6, 679–684.
- Berlyne, D. E. (1966). Curiosity and exploration. *Science* 153, 25–33. doi: 10.1126/science.153.3731.25
- Bouguet, J. Y. (2009). *Camera Calibration Toolbox for Matlab*. Available online at: <http://www.vision.caltech.edu/bouguetj/>
- Brafman, R., and Tennenholtz, M. (2003). R-max—a general polynomial time algorithm for near-optimal reinforcement learning. *J. Mach. Learn. Res.* 3, 213–231. doi: 10.1162/153244303765208377
- Butko, N. J., and Movellan, J. R. (2010). Infomax control of eye movements. *IEEE Trans. Auton. Ment. Dev.* 2, 91–107. doi: 10.1109/TAMD.2010.2051029
- Cavallanti, G., Cesa-bianchi, N., and Gentile, C. (2008). Linear classification and selective sampling under low noise conditions. *Adv. Neural Inform. Process. Syst.* 21, 249–256.
- Cesa-Bianchi, N., Conconi, A., and Gentile, C. (2005). A second-order perceptron algorithm. *SIAM J. Comput.* 34, 640–668. doi: 10.1137/S0097539703432542
- Cesa-Bianchi, N., Gentile, C., and Orabona, F. (2009). “Robust bounds for classification via selective sampling,” in *Proceedings of International Conference on Machine Learning (ICML)*, (Montreal), 121–128.
- Cesa-Bianchi, N., and Lugosi, G. (2006). *Prediction, learning, and games*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511546921
- Dekel, O., Gentile, C., and Sridharan, K. (2010). “Robust selective sampling from single and multiple teachers,” in *Proceedings of Conference on Learning Theory (COLT)*, (Haifa, IL), 346–358.
- Dekel, O., Gentile, C., and Sridharan, K. (2012). Selective sampling and active learning from single and multiple teachers. *J. Mach. Learn. Res.* 13, 2655–2697.
- Freund, Y., Seung, H. S., Shamir, E., and Tishby, N. (1997). Selective sampling using the query by committee algorithm. *Mach. Learn.* 28, 133–168. doi: 10.1023/A:1007330508534
- Gordon, G., and Ahissar, E. (2011). “Reinforcement active learning hierarchical loops,” in *Proceedings of International Joint Conference on Neural Networks (IJCNN)*, (San Jose, CA: IEEE), 3008–3015.
- Hester, T., and Stone, P. (2012). “Intrinsically motivated model learning for a developing curious agent,” in *Proceedings of International Conference on Development and Learning and Epigenetic Robotics (ICDL-EPIROB)*, (San Diego, CA). doi: 10.1109/DevLrn.2012.6400082
- Howard, R. A. (1960). *Dynamic programming and markov processes*. Cambridge, MA: The MIT Press.
- Kearns, M., and Singh, S. (2002). Near-optimal reinforcement learning in polynomial time. *Mach. Learn.* 49, 209–232. doi: 10.1023/A:1017984413808
- Kocsis, L., and Szepesvári, C. (2006). “Bandit based monte-carlo planning,” in *Proceedings of European Conference on Machine Learning (ECML)*, (Berlin; Heidelberg), 282–293.
- Konidaris, G. (2011). *Autonomous robot skill acquisition*. PhD thesis, University of Massachusetts Amherst.
- Lagoudakis, M. G., and Parr, R. (2003). Least-squares policy iteration. *J. Mach. Learn. Res.* 4, 1107–1149. doi: 10.1162/jmlr.2003.4.6.1107
- Lang, T. (2011). *Planning and Exploration in Stochastic Relational Worlds*. PhD thesis, Fachbereich Mathematik und Informatik, Freie Universität Berlin.
- Lang, T., and Toussaint, M. (2009). “Relevance grounding for planning in relational domains,” in *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases: Part I*, (Springer-Verlag), 736–751.
- Lopes, M., Lang, T., Toussaint, M., and Oudeyer, P.-Y. (2012). “Exploration in model-based reinforcement learning by empirically estimating learning progress” in *Neural Information Processing Systems (NIPS)*, (Tahoe).
- Luciw, M., Graziano, V., Ring, M., and Schmidhuber, J. (2011). “Artificial curiosity with planning for autonomous perceptual and cognitive development,” in *Proceedings of International Conference on Development and Learning and Epigenetic Robotics (ICDL-EPIROB)*, (Frankfurt).
- Lungarella, M., Metta, G., Pfeifer, R., and Sandini, G. (2003). Developmental robotics: a survey. *Connect. Sci.* 15, 151–190. doi: 10.1080/09540090310001655110
- Mausam and Kolobov, A. (2012). *Planning with Markov Decision Processes: An AI Perspective*. (*Synthesis Lectures on Artificial Intelligence and Machine Learning*). San Rafael, CA: Morgan & Claypool Publishers.
- Meuleau, N., and Bourgine, P. (1999). Exploration of multi-state environments: local measures and back-propagation of uncertainty. *Mach. Learn.* 35, 117–154. doi: 10.1023/A:1007541107674
- Neuronics, A. G. (2004). *Katana User Manual and Technical Description*. Available online at: <http://www.neuronics.ch/>
- Newell, A., Shaw, J. C., and Simon, H. A. (1959). “Report on a general problem-solving program,” in *IFIP Congress*, 256–264.
- Ngo, H., Luciw, M., Förster, A., and Schmidhuber, J. (2012). “Learning skills from play: artificial curiosity on a katana robot arm,” in *Proceedings of International Joint Conference of Neural Networks (IJCNN)*, (Brisbane, QLD), 1–8.
- Ngo, H., Ring, M., and Schmidhuber, J. (2011). “Compression progress-based curiosity drive for developmental learning,” in *Proceedings of the 2011 IEEE*

- Conference on Development and Learning and Epigenetic Robotics IEEE-ICDL-EPIROB.* (IEEE).
- Orabona, F., and Cesa-Bianchi, N. (2011). "Better algorithms for selective sampling," in *Proceedings of International Conference on Machine Learning (ICML)*, (Washington, DC), 433–440.
- Oudeyer, P.-Y., Kaplan, F., and Hafner, V. V. (2007). Intrinsic motivation systems for autonomous mental development. *IEEE Trans. Evol. Comput.* 11, 265–286. doi: 10.1109/TEVC.2006.890271
- Piaget, J. (1955). *The Child's Construction of Reality*. London: Routledge and Kegan Paul.
- Rao, K., and Whiteson, S. (2012). "V-max: tempered optimism for better pac reinforcement learning," in *Proceedings of Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, (Valencia: International Foundation for Autonomous Agents and Multiagent Systems), 375–382.
- Ring, M. B. (1994). *Continual Learning in Reinforcement Environments*. PhD thesis, University of Texas, (Austin, TX), 78712.
- Ring, M. B. (1997). Child: a first step towards continual learning. *Mach. Learn.* 28, 77–104. doi: 10.1023/A:1007331723572
- Schmidhuber, J. (1991). "Curious model-building control systems," in *Proceedings of International Joint Conference on Neural Networks (IJCNN)*, (Singapore: IEEE), 1458–1463.
- Schmidhuber, J. (1997). *What's interesting?* Technical Report IDSIA-35-97, IDSIA. Available online at: <ftp://ftp.idsia.ch/pub/juergen/interest.ps.gz>; extended abstract in Proceedings of Snowbird'98, Utah, 1998; see also Schmidhuber (2002).
- Schmidhuber, J. (2002). "Exploring the predictable," in *Advances in Evolutionary Computing*, eds A. Ghosh and S. Tsuitsui (Berlin; Heidelberg: Springer), 579–612.
- Schmidhuber, J. (2006). Developmental robotics, optimal artificial curiosity, creativity, music, and the fine arts. *Connect. Sci.* 18, 173–187. doi: 10.1080/09540090600768658
- Schmidhuber, J. (2010). Formal theory of creativity, fun, and intrinsic motivation (1990–2010). *IEEE Trans. Auton. Ment. Dev.* 2, 230–247. doi: 10.1109/TAMD.2010.2056368
- Schmidhuber, J., and Huber, R. (1991). Learning to generate artificial fovea trajectories for target detection. *Int. J. Neural Syst.* 2, 135–141. doi: 10.1142/S012906579100001X
- Şimşek, Ö., and Barto, A. G. (2006). "An intrinsic reward mechanism for efficient exploration," in *Proceedings of International Conference on Machine Learning (ICML)*, (ACM), 833–840.
- Srivastava, R. K., Steunebrink, B. R., and Schmidhuber, J. (2013). First Experiments with POWERPLAY. *Neural Netw.* 41, 130–136. doi: 10.1016/j.neunet.2013.01.022
- Steinwart, I. (2002). On the influence of the kernel on the consistency of support vector machines. *J. Mach. Learn. Res.* 2, 67–93. doi: 10.1162/153244302760185252
- Storck, J., Hochreiter, S., and Schmidhuber, J. (1995). "Reinforcement driven information acquisition in non-deterministic environments," in *Proceedings of International Conference on Artificial Neural Networks (ICANN)*, Vol. 2, (Paris), 159–164.
- Strehl, A., and Littman, M. (2008). Online linear regression and its application to model-based reinforcement learning. *Adv. Neural Inform. Process. Syst.* 20, 1417–1424. doi: 10.1016/j.jcoss.2007.08.009
- Sutton, R., and Barto, A. (1998). *Reinforcement Learning: An Introduction*, Vol. 28. Cambridge, MA: MIT press.
- Sutton, R., Modayil, J., Delp, M., Degris, T., Pilarski, P., White, A., et al. (2011). "Horde: a scalable real-time architecture for learning knowledge from unsupervised sensorimotor interaction," in *Proceedings of International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, (Taipei), 761–768.
- Sutton, R. S. (1990). "Integrated architectures for learning, planning, and reacting based on approximating dynamic programming," in *Proceeding International Conference on Machine Learning (ICML)*, 216–224.
- Szepesvári, C. (2010). "Algorithms for reinforcement learning," in *Synthesis Lectures on Artificial Intelligence and Machine Learning*, Vol. 4, (San Rafael, CA), 1–103.
- Thrun, S., and Mitchell, T. (1995). Lifelong robot learning. *Robot. Autonom. Syst.* 15, 25–46. doi: 10.1016/0921-8890(95)00004-Y
- Vovk, V. (2001). Competitive on-line statistics. *Int. Stat. Rev.* 69, 213–248. doi: 10.1111/j.1751-5823.2001.tb00457.x
- Watkins, C. J. C. H., and Dayan, P. (1992). Q-learning. *Mach. Learn.* 8, 279–292. doi: 10.1007/BF00992698
- Weng, J., McClelland, J., Pentland, A., Sporns, O., Stockman, I., Sur, M., et al. (2001). Autonomous mental development by robots and animals. *Science* 291, 599–600. doi: 10.1126/science.291.5504.599
- Whitehead, S. D., and Ballard, D. H. (1990). Active perception and reinforcement learning. *Neural Comput.* 2, 409–419. doi: 10.1162/neco.1990.2.4.409
- Wiering, M., and Schmidhuber, J. (1998). "Efficient model-based exploration," in *Proceedings of International Conference on Simulation of Adaptive Behaviour (SAB)*, (Zurich), 223–228.
- Yi, S., Gomez, F., and Schmidhuber, J. (2011). "Planning to be surprised: optimal Bayesian exploration in dynamic environments," in *Proceedings of Fourth Conference on Artificial General Intelligence (AGI)*, (Mountain View, CA: Google).

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 12 July 2013; accepted: 21 October 2013; published online: 26 November 2013.

*Citation: Ngo H, Luciw M, Förster A and Schmidhuber J (2013) Confidence-based progress-driven self-generated goals for skill acquisition in developmental robots. *Front. Psychol.* 4:833. doi: 10.3389/fpsyg.2013.00833*

This article was submitted to Cognitive Science, a section of the journal Frontiers in Psychology.

Copyright © 2013 Ngo, Luciw, Förster and Schmidhuber. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Linear combination of one-step predictive information with an external reward in an episodic policy gradient setting: a critical analysis

Keyan Zahedi^{1*}, Georg Martius¹ and Nihat Ay^{1,2}

¹ Max Planck Institute for Mathematics in the Sciences, Leipzig, Germany

² Santa Fe Institute, Santa Fe, NM, USA

Edited by:

Gianluca Baldassarre, Italian National Research Council, Italy

Reviewed by:

Jürgen Schmidhuber, Istituto Dalle Molle di Studi sull'Intelligenza Artificiale, Switzerland

Chrisantha T. Fernando, University of Sussex, UK

Matthew D. Luciw, Istituto Dalle Molle di Studi sull'Intelligenza Artificiale, Switzerland

***Correspondence:**

Keyan Zahedi, Max Planck Institute for Mathematics in the Sciences, Inselstrasse 22, 04103 Leipzig, Germany

e-mail: zahedi@mis.mpg.de

One of the main challenges in the field of embodied artificial intelligence is the open-ended autonomous learning of complex behaviors. Our approach is to use task-independent, information-driven intrinsic motivation(s) to support task-dependent learning. The work presented here is a preliminary step in which we investigate the predictive information (the mutual information of the past and future of the sensor stream) as an intrinsic drive, ideally supporting any kind of task acquisition. Previous experiments have shown that the predictive information (PI) is a good candidate to support autonomous, open-ended learning of complex behaviors, because a maximization of the PI corresponds to an exploration of morphology- and environment-dependent behavioral regularities. The idea is that these regularities can then be exploited in order to solve any given task. Three different experiments are presented and their results lead to the conclusion that the linear combination of the one-step PI with an external reward function is not generally recommended in an episodic policy gradient setting. Only for hard tasks a great speed-up can be achieved at the cost of an asymptotic performance lost.

Keywords: information-driven self-organization, predictive information, reinforcement learning, embodied artificial intelligence, embodied machine learning

1. INTRODUCTION

One of the main challenges in the field of embodied artificial intelligence (EAI) is the open-ended autonomous learning of complex behaviors. Our approach is to use task-independent, information-driven intrinsic motivation to support task-dependent learning in the context of reinforcement learning (RL) and EAI. The work presented here is a first step into this direction. RL is of growing importance in the field of EAI, mainly for two reasons. First, it allows to learn the behaviors of high-dimensional and complex systems with simple objective functions. Second, it has a well-established theoretical (Sutton and Barto, 1998; Bellman, 2003) and biological foundation (Dayan and Balleine, 2002). In the context of EAI, where the agent has a morphology and is situated in an environment, the concepts of the agent's intrinsic and extrinsic perspective rise naturally. As a direct consequence, several questions about intrinsic and extrinsic reward functions, denoted by IRF and ERF, follow from the EAI's point of view. The questions that are of interest to us are; what distinguishes an IRF from an ERF, what is a good candidate for a first principled IRF, and finally, how should IRFs and ERFs be combined?

The first question, of how to distinguish between IRF and ERF is addressed in the second section of this work, which starts with the conceptual framework of the sensorimotor loop and its representation as a causal graph. This leads to a natural distinction of variables that are intrinsic and extrinsic to the agent. We define an IRF that models an internal drive or motivation as a

task-independent function which operates on the agent's intrinsic variables only. In general, an ERF is a task-dependent function that may operate on intrinsic and extrinsic variables.

The main focus of this work is the second question, which deals with finding a first principled IRF. We propose the predictive information (PI) (Bialek et al., 2001) for the following reasons. Information-driven self-organization, by the means of maximizing the one-step approximation of the PI has proved to produce a coordinated behavior among physically coupled but otherwise independent agents (Ay et al., 2008; Zahedi et al., 2010). The reason is that the PI inherently addresses two important issues of self-organized adaptation, as the following equation shows: $I(S_t; S_{t+1}) = H(S_{t+1}) - H(S_{t+1}|S_t)$, where S_t is the vector of intrinsically accessible sensor values at time t . The first term leads to a diversity of the behavior, as every possible sensor state must be visited with equal probability. The second term ensures that the behavior is compliant with the constraints given by the environment and the morphology, as the behavior must be predictable. This means that an agent maximizing the PI explores behavioral regularities, which can then be exploited to solve a task. In a differently motivated work, namely to obtain purely self-organizing behavior, a time-local version of the PI was successfully used to drive the learning process of a robot controller (Martius et al., 2013). A similar learning rule was obtained from the principle of Homeokinesis (Der and Martius, 2012). In both cases a gradient information was derived to pursue local optimization. For the integration of external goals a set of methods have been proposed

by (Martius and Herrmann, 2012), which, however, cannot deal with the standard reinforcement setting of arbitrary time-delayed rewards that we study here. Prokopenko et al. (2006) used the PI, estimated on the spatio-temporal phase-space of an embodied system, as part of fitness function in an artificial evolution setting. It was shown that the resulting locomotion behavior of a snake-bot was more robust, compared to the setting, in which only the traveled distance determined the fitness.

The third question, which deals with how to combine the IRF and ERF, is in the focus of the ongoing research that was briefly described above and of which this publication is a first step. As the PI maximization is considered to be an exploration of behavioral regularities, it would be natural to exchange the exploration method of a RL algorithm by a gradient on the PI. The work presented here is a preliminary step in which we concentrate on the effect of the PI in a RL context to understand for which type of learning problems it is beneficial and in which it might not be. Therefore, we chose a linear combination of IRF and ERF in an episodic RL setting to evaluate the PI as an IRF in different experiments. Combining an IRF and an ERF in this way is justified as ERFs are often linear combinations of different terms, such as one term for fast locomotion and another for low energy consumption. Nevertheless, the results of the experiments presented in this work show that the one-step PI should not be combined in this way with an ERF in an episodic policy gradient setting.

We are not the first to address the question of IRF and ERF in the context of RL and EAI. This idea goes back to the pioneering work of Schmidhuber (1990) and is also in the focus of more recent work (Kaplan and Oudeyer, 2004; Schmidhuber, 2006; Oudeyer et al., 2007) which are based on the prediction progress and Barto et al. (2004), who considers the prediction error. In Storck et al. (1995); Yi et al. (2011) an intrinsic reward for information gain was proposed (KL-divergence between subsequent models), which results in their experiments in a state-entropy maximization. A different approach (Little and Sommer, 2013) uses a greedy policy on the predicted information gain of the world model to select the next action of an agent. However, only discrete state/action spaces have been considered in both approaches. A similar work (Cuccu et al., 2011) uses compression quality as the intrinsic motivation, which was particularly beneficial because it performed a reduction of the high-dimensional visual input space. In comparison to our work only one experiment (comparable to the self-rescue task below) with a one-dimensional action-space was used without considering asymptotic performance, which is where we found most problems.

This paper investigates continuous space high-dimensional control problems where random exploration becomes difficult. The PI, measured on the sensor values, accompanies (and might eventually replace) the exploration of a RL method such that the policy adaptations are conducted compliant to the morphology and environment. The actual embodiment is taken into account, without modeling it explicitly in the learning process.

The work is organized in the following way. The next section gives an overview of the methods, beginning with the sensorimotor loop and its causal representation. This is then followed by a presentation of the PI and the episodic RL method PGPE (Sehnke

et al., 2010). The third section describes the results received by applying the methods to three experiments, and the last section closes with a discussion.

2. METHODS

This section describes the methods used in this work. It begins with the conceptual framework of the sensorimotor loop. This is then followed by a discussion of the PI and entropy, which both are used as IRF in all presented experiments. Finally, the RL algorithm utilized in this work is introduced as far as it is required to understand how the results were obtained.

2.1. EMBODIED AGENTS AND THE SENSORIMOTOR LOOP

There are three main reasons why we prefer to experiment with embodied agents (EA). First, *scalability*: EA are high-dimensional systems which live in a continuous world. Hence, the algorithms face the curse of dimensionality if they are evaluated on different EAs. Second, *validation*: we are interested in understanding natural cognitive systems by the means of building artificial agents (Brooks, 1991). Using EA ensures that the models are validated against the same (or similar) physical constraints that natural systems are exposed to. Third, *guidance*: there is good evidence that the constraints posed by the morphology and environment can be used to reduce the required controller complexity, and hence, reduce the size of the search space for a learning algorithm (Zahedi et al., 2010; Pfeifer and Bongard, 2006). Consequently, understanding the interplay between the body, brain and environment, also called the sensorimotor loop (SML, see **Figure 1**), is a general focus of our work. The next paragraph will introduce the general concept of the SML and discuss its representation as a causal graph.

A cognitive system consists of a brain or controller that sends signals to the system's actuators, which then affect the system's environment. We prefer the notion of the system's *Umwelt* (von Uexküll, 1934; Clark, 1996; Zahedi et al., 2010; Zahedi and Ay, 2013), which is the part of the system's environment that can be affected by the system, and which itself affects the system. The state of the actuators and the *Umwelt* are not directly accessible to the cognitive system, but the loop is closed as information about both, the *Umwelt* and the actuators are provided to the controller by the system's sensors. In addition to this general concept, which is widely used in the EAI community (see e.g., Pfeifer et al., 2007), we introduce the notion of *world* to the sensorimotor loop, and by that we mean the system's morphology and the system's *Umwelt*. We can now distinguish between the agent's intrinsic and extrinsic perspective in this context. The world is everything that is extrinsic from the perspective of the cognitive system, whereas the controller, sensor and actuator signals are intrinsic to the system.

The distinction between intrinsic and extrinsic is also captured in the representation of the sensorimotor loop as a causal or Bayesian graph (see **Figure 1**, right-hand side). The random variables C , A , W , and S refer to the controller state, actuator signals, world and sensor signals, and the directed edges reflect causal dependencies between the random variables (see Klyubin et al., 2004; Ay and Polani, 2008; Zahedi et al., 2010). Everything that is extrinsic to the system is captured in the variable W , whereas S , C , and A are intrinsic to the system.

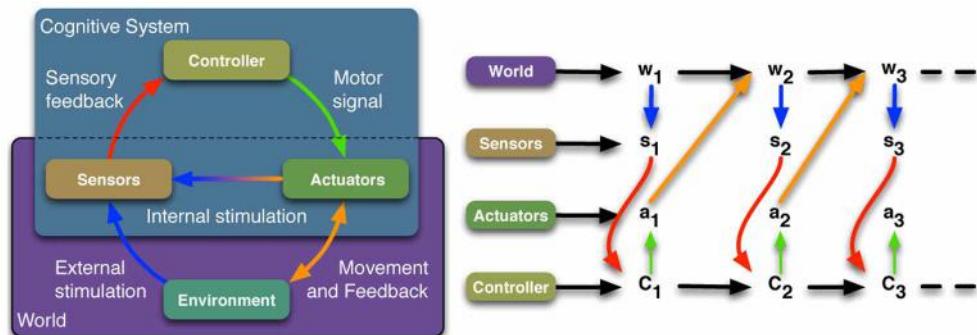


FIGURE 1 | The sensorimotor loop. Left: Schematic diagram of a cognitive system with its interaction with the world. **Right:** Corresponding causal graph.

In this context, we distinguish between internal and external reward function (IRF, ERF) in the following way. An ERF may access any variable, especially those that are not available to an agent by its sensors, i.e., anything that we summarized as the world state W . An IRF may access intrinsically available information only (S_t, A_t, C_t , see Figure 1). We are interested in first principled model of an intrinsic motivation, i.e., a model that requires as few assumptions as possible. The idea is that IRF should not depend on a specific task but rather be a task-independent internal driving force, which supports any task-dependent learning. This is why we refer to it as task-independent internal motivation or drive. This closes the discussion of embodied agents and their formalization in terms of the sensorimotor loop. The next section describes the information-theoretic measures that are used in the remainder of this work.

2.2. PREDICTIVE INFORMATION

The predictive information (PI) (Bialek et al., 2001), which is also known as excess entropy (Crutchfield and Young, 1989) and effective measure complexity (Grassberger, 1986) is defined as the mutual information of the entire past and future of the sensor data stream:

$$I_{\text{pred}}(S) := I(S_p; S_f) \quad (1)$$

where $S_p = \{S_1, S_2, \dots, S_t\}$ is the entire past of the system's sensor data at some time $t \in \mathbb{N}$ and $S_f = \{S_{t+1}, S_{t+2}, \dots\}$ its entire future. The PI captures how much information the past carries about the future. Unfortunately, it cannot be calculated for most applications because of technical reasons. Hence, we use the one-step PI, which is given by

$$\begin{aligned} I_{\text{pred}}^*(S) &:= I(S_{t+1}; S_t) \\ &= \underbrace{H(S_{t+1})}_{\text{diversity}} - \underbrace{H(S_{t+1}|S_t)}_{\text{compliance}}, \end{aligned} \quad (2)$$

which was previously investigated in the context of EAI (Ay et al., 2008) and as a first principle learning rule (Zahedi et al., 2010; Martius et al., 2013). A different motivation for the PI

is based on maximizing the mutual information of an intention state \tilde{S}_t , which is internally generated by the agent, and the next sensor state S_{t+1} (Ay and Zahedi, 2013). The Equation (2) displays how maximizing the PI affects the behavior of a system. The first term in Equation (2) leads to a maximization of the entropy over the sensor states. This means that the agent has to explore its world in order to sense every state with equal probability. The second term in Equation (2) states that the uncertainty of the next sensor state must be minimal if the current sensor state is known. This means that an agent has to choose actions which lead to predictable next sensor states. This can be rephrased in the following way. Maximizing the entropy $H(S_{t+1})$ increases the diversity of the behavior whereas minimizing the conditional entropy $-H(S_{t+1}|S_t)$ increases the compliance of the behavior. The result is a system that explores its sensors space to find as many regularities in its behavior as possible.

For completeness we will also maximize the entropy $H(S_t)$ only and compare the results to the maximization of the PI. This concludes the presentation of the PI (and entropy) as a model for a task-independent internal motivation in the context of RL. The next section presents the utilized RL algorithm.

2.3. POLICY GRADIENTS WITH PARAMETER-BASED EXPLORATION (PGPE)

We chose an episodic RL method named PGPE (Sehnke et al., 2010) to investigate the effect of the PI as an IRF, because it is not restricted to a specific class of policies. Any policy, which can be represented by a vector $\mu \in \mathbb{R}^n$ with fixed length $n \in \mathbb{N}^+$ can be optimized by this method. In the work presented here, we use it to learn the synaptic strengths and bias values of neural networks with fixed structures only. Nevertheless, we can apply the framework to other parametrizations, in particular to stochastic policies, which is why PGPE attracted our attention for ongoing the project in which this work is embedded.

The algorithm can be summarized in the following way (for details, see (Sehnke et al., 2010)). In each *roll-out* or episode, two policy instances are drawn from μ by adding and subtracting a random vector $\epsilon \sim \mathcal{N}(0, \sigma)$ to it. The resulting two policy parametrizations $\Theta^+ = \mu + \epsilon$ and $\Theta^- = \mu - \epsilon$ are then

evaluated and their final rewards r^+, r^- are used to determine the modifications on μ and σ according to the following equations

$$m^n = \max(m^{n-1}, r^{+, n}, r^{-, n}) \quad (3)$$

$$b^n = (1 - \delta)b^{n-1} + \delta \sum_n \frac{r^{+, n} + r^{-, n}}{2} \quad (4)$$

$$\Delta\mu_i = \frac{\alpha\epsilon_i(r^+ - r^-)}{2m - r^+ - r^-} \quad (5)$$

$$\Delta\sigma_i = \frac{\alpha}{m - b} \left(\frac{r^+ - r^-}{2} - b \right) \left(\frac{\epsilon^2 - \sigma_i^2}{\sigma_i} \right). \quad (6)$$

Roll-outs can be repeated several times before a learning step is performed. Every learning step concludes a *batch*. PGPE requires an initial μ_{init} , an initial σ_{init} , a learning rate α , baseline b , baseline adaptation parameter δ , and an initialized maximal reward $m = m_{\text{init}}$. We have set δ to the recommended value of 0.1, $\mu_{\text{init}} = 0$, and we have achieved the best results in all experiments by setting m_{init} small enough that m is definitely overwritten in the first roll-out (see Equation (3)). The other parameters are evaluated in each experiment, such that the best results were achieved when no IRF was used and then fixed for the remaining experiments.

3. RESULTS

This section presents three different experiments and their results. The first experiment is the cart-pole swing-up, a standard control theory problem that is also widely used in machine learning (Barto et al., 1983; Geva and Sitte, 1993; Doya, 2000; Pasemann et al., 1999). The cart-pole experiment is also chosen because balancing a pole minimizes the entropy, and hence, it contradicts the maximization of the PI. The second experiment is the learning of a locomotion behavior for a hexapod and it was chosen to demonstrate the effect of the PI maximization on a more common, well-structured experimental setting. By well-structured we mean that the controller, morphology, environment, and ERF are chosen such that they result in a good hexapod locomotion without any additional support by an IRF in only a few policy updates. The third experiment is designed to be challenging, as it combines a high-dimensional system, an unconventional control structure, an unsteady ERF with an unsteady environment. We believe that these three experiments span a broad range of possible applications for information-theoretic IRF in the context of episodic RL.

3.1. CART-POLE SWING-UP

The cart-pole swing-up experiment is ideal to investigate the effect of the PI on an episodic RL task, mainly for two reasons. First, the experiment is well-defined by a set of equations and parameters that are widely used in literature (Barto et al., 1983; Geva and Sitte, 1993; Doya, 2000; Pasemann et al., 1999). This ensures that the results are comparable and reproducible by others with little effort. Second, the successful execution of the task contradicts the maximization of the PI. The task is to balance the pole in the center of the environment, and hence, to minimize the entropy of the sensor states. The maximization of the PI demands

a maximization of the entropy (see Equation 2). The remainder of this section first describes the experimental and controller setting and then closes with a discussion of the results.

The experiment was conducted by implementing the equations that can be found in (Barto et al., 1983; Geva and Sitte, 1993; Doya, 2000). The state of the cart-pole is given by $x, \dot{x}, \vartheta, \dot{\vartheta}$, which are the position of the cart, the speed of the cart, the pole angle and the pole's angular velocity. The cart is controlled by a force $F \in [-10N, 10N]$ that is applied to its center of mass. The four state variables and the force define the input and output configuration of our controllers for this task. The initial controller (see Figure 2A) was chosen from (Pasemann et al., 1999), where network structures were evolved for the same task. To ensure that the evolved structure is not especially unsuitable for RL, different variations were chosen for evaluation too (see Figures 2B–D). In this approach, the input neurons are simple buffer neurons, with the identity as transfer-function, whereas all other neurons use the hyperbolic tangent transfer-function.

The evaluation time was set to $T = 2000$ iterations, which corresponds to 20 seconds (c.f. Doya, 2000). Different values, starting from the values proposed in (Sehnke et al., 2010), for the learning rate $\alpha \in \{0.1, 0.2, 0.5\}$, the initial variation $\sigma_{\text{init}} \in \{2, 5\}$, and the initial maximal reward $m_{\text{init}} \in \{-\infty, 10, 100, 1000\}$ were evaluated in experiments without applying an IRF to the learning of the task. The underlined values showed the best results, and hence, are chosen for presentation here. Each experiment consisted of $B = 10000$ batches, i.e., updates of μ and σ (see Equations 5 and 6) with two roll-outs each (i.e., four evaluated policies $\theta_{1,2}^{+, -}$). The results are obtained by conducting every experiment 100 times. To ensure comparability among the experiments with different parameters and controllers, the random number generator was initialized from a fixed set of 100 integer values for each experiment.

The presentation of the reward function is split into two parts. The first part handles the ERF, whereas the second part handles the IRF. We use the terms *intrinsic/internal* and *extrinsic/external* with respect to the agent's perspective, as discussed in the previous section (see Section 2.1). The controller has access to the full state of the system, and hence, the separation into internal and external is artificial in this case. Nevertheless, we keep this terminology for consistency, as the next experiments will reflect this distinction in a natural way. We denote IRF by R_{in} and ERF by R_{ex} , where a super-script is added to distinguish between the different reward functions (PI and entropy).

The ERF for the cart-pole swing-up task is defined such that it is not a smooth gradient in the reward space, and therefore, does not directly guide the learning process. The controller is only rewarded if the pole is pointing upwards and the reward is scaled with the distance of the pole to the center of the environment, which is given by

$$R_{\text{ex}}(t) := \begin{cases} 2 - |x(t)| & \text{if } |\vartheta(t)| < 5^\circ \\ 0 & \text{otherwise.} \end{cases} \quad (7)$$

The IRF is calculated at the end of each episode based on the recordings of the pole angles $\{S_t = \vartheta(t) | t = 1, 2, \dots, T\}$.

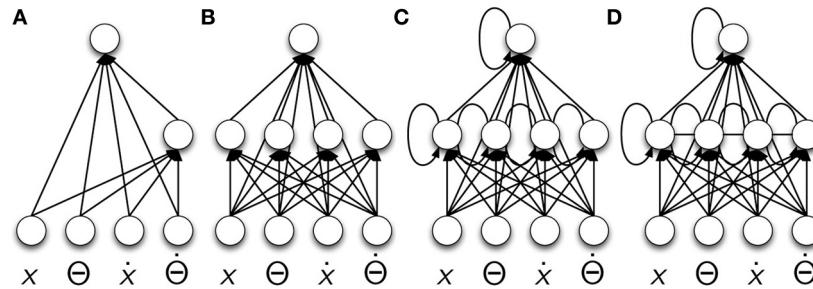


FIGURE 2 | Controller architectures for the cart-pole swing-up task. The input neurons are bare buffer neurons whereas the hidden and output neurons have tanh transfer-function. **(A)** from

Pasemann et al. (1999); **(B)** with 4 hidden neurons and fully connected; **(C,D)** recurrent variations without and with lateral connections.

We use a discrete-valued computation of the PI, and hence, the data is binned prior to the calculation. All IRFs are normalized with respect to their theoretical upper bound of $I(S_{t+1}; S_t) \leq H(S_t) \leq \log |S|$ (see (Cover and Thomas, 2006)). This leads to the two following IRFs:

$$R_{\text{in}}^{\text{PI}} := |I(S_{t+1}; S_t)| \quad \text{and} \quad R_{\text{in}}^{\text{H}} := |H(S_t)|. \quad (8)$$

The overall reward functions are then given by

$$\begin{aligned} R^{\text{PI}} &:= \sum_{t=1}^T R_{\text{ex}}(t) + \beta(\gamma) R_{\text{in}}^{\text{PI}}, \\ R^{\text{H}} &:= \sum_{t=1}^T R_{\text{ex}}(t) + \beta(\gamma) R_{\text{in}}^{\text{H}}, \quad \beta(\gamma) = \gamma \cdot T \cdot \max_{x, \vartheta, t} \{R_{\text{ex}}(t)\} \end{aligned} \quad (9)$$

where $\beta(\gamma)$ is a factor to scale the IRF with respect to the maximal possible value of the ERF. This allows us to compare the effects of $R_{\text{in}}^{\text{PI}}$ and R_{in}^{H} across different experiments.

The results are discussed only for the fully connect feed-forward network (see **Figures 3A–D**) in detail as this controller shows the most distinguishable results with respect to the variation of the IRF scaling parameter $\gamma \in \{0, 1.25, 2.5, 3.75, \text{ and } 5\%\}$. It is important to note that the plots only show the averages of the 100 experiments and not the standard deviation for the following reason. Few controllers succeed early, others later during the process. Due to the unsteady ERF the resulting standard deviation is very large, as those controllers that succeed receive significantly higher reward compared to those not succeeding (which remain close to zero, as a rotational behavior is not permitted). We intentionally chose an unsteady ERF, that returns zero for almost all states, and hence, we know beforehand that the standard deviation is large and no further information is provided if it is plotted.

Figures 3A,B show the progress of the ERF $R_{\text{ex}}^{\text{PI}}$ and IRF $R_{\text{in}}^{\text{PI}}$ for the PI maximization. It is shown that there is a significant speed-up in learning during the first 4000 batches for all $\gamma > 0\%$ (see **Figure 3A**). At this point in time the average ERF of $\gamma = 0\%$ succeeds that of $\gamma = 5\%$. After approximately 5000 batches the ERF for $\gamma = 2.5\%$ and $\gamma = 3.75\%$ are very close to or slightly succeeded by the ERF for $\gamma = 0\%$, whereas the ERF for $\gamma =$

1.25% remains higher. The conclusion from this experiment is that small values of $\gamma < 5\%$ are beneficial in this learning task as less batches are required to solve this task and the asymptotic learning performances are almost identical to $\gamma = 0\%$. The results, however, are not significant and the choice of γ is critical. This leads to the conclusion that the one-step PI is not significantly beneficial in the learning of this task.

Figures 3C,D show the progress of the ERF R_{ex}^{H} and IRF R_{in}^{H} for the entropy maximization. The results show a different picture. Any parameter $\gamma > 0\%$ speeds up the learning and improves the overall performance. The comparison of entropy and PI is addressed in the discussion again.

3.2. HEXAPOD LOCOMOTION

If a specific task should be learned by an embodied agent, it is more common to choose an environment, morphology, control structure and a smooth ERF which are well-suited for the desired task. In order to investigate which effect the PI has on such a well-defined learning task, the set-up of the experiment presented in this section is chosen such that all components are known to work well if there is no IRF present. The goal is to learn a locomotion behavior of a hexapod, where the maximal deviation angles ensure that it cannot flip over. The controller is known to perform well in a similar task (Markelić and Zahedi, 2007) and its modularity significantly reduces the number of parameters that must be learned. The ERF defines a smooth gradient in the reward space, ensuring that small changes in the controller parameters show an immediate effect in the ERF. The environment is an even plane without any obstacles.

The experimental platform (see **Figure 4**) is a hexapod, with 12 degrees of freedom (two actuators in each leg) and with 18 sensors (angular positions of the actuators and binary foot contact sensors). The two actuators of each leg are positioned in the shoulder (Thorax-Coxa or ThC joint) and in the knee (Femur-Tibia or FTi joint) of the walking machine, similar to the morphology presented in (von Twickel et al., 2011). We omit the second shoulder-joint (CTr) because it is not required for locomotion. Each joint accepts the desired angular position as its input and returns the actual current angular position as its output. The simulator YARS (Zahedi et al., 2008) was used for all experiments conducted in this section.

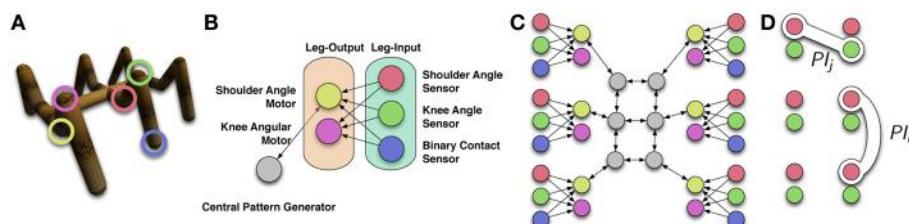
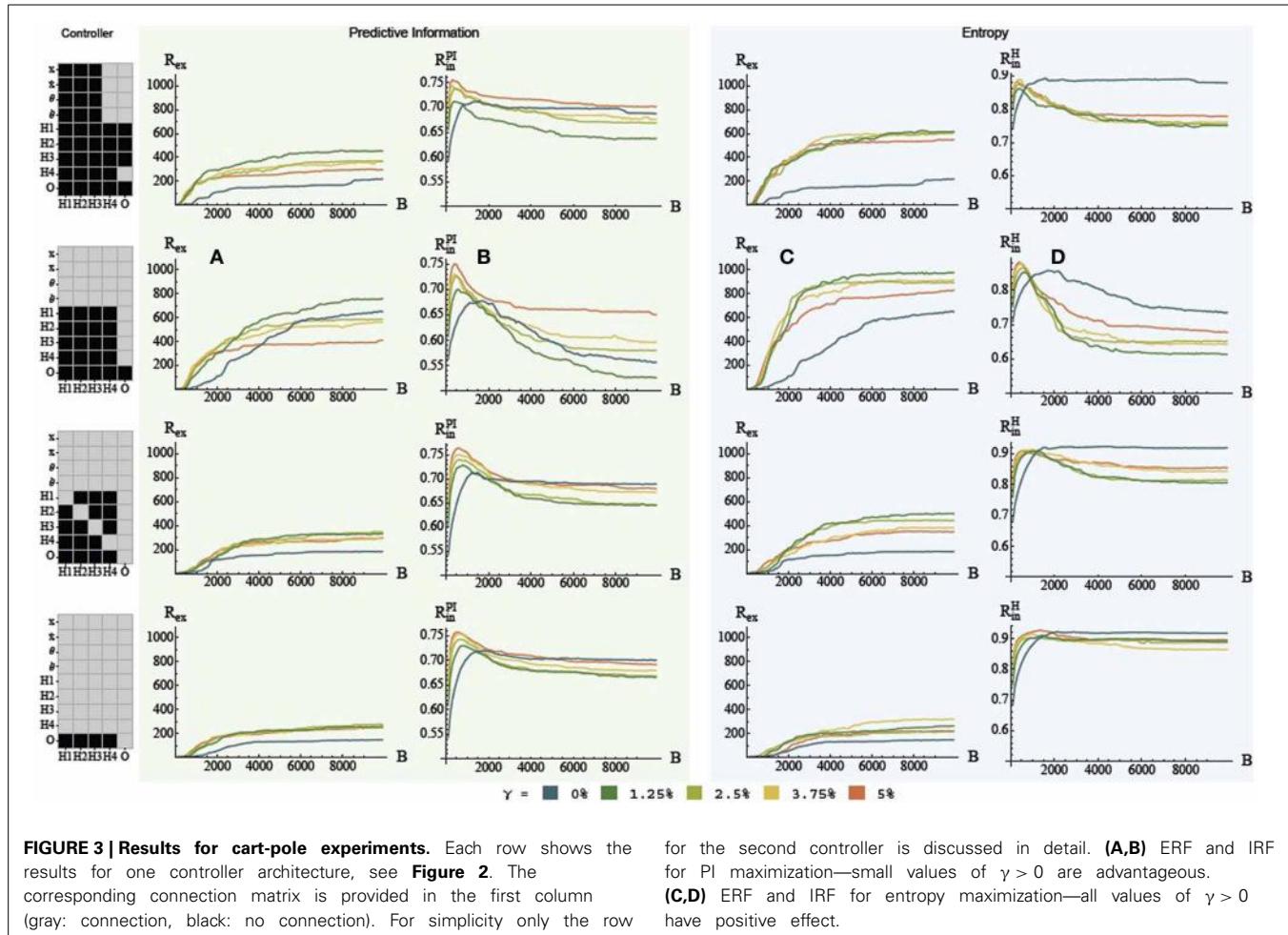


FIGURE 4 | Hexapod for locomotion task and controller set-up. **(A)** Hexapod robot with marked actuated joints and sensors; **(B)** leg module of controller; **(C)** entire controller; and **(D)** schematic pairings for PI and entropy calculation.

Different values for the PGPE parameters were evaluated. The best results for $\gamma = 0$ (see Equation 9) were achieved with $\sigma_{\text{init}} = 2$ and $\alpha = 0.1$. To ensure comparability with the previous experiment, two roll-outs were chosen here, although it is not required to obtain the following results. The evaluation time was set to $T = 1000$ and $B = 250$ batches were sufficient to observe a convergence of the policy parameters μ . The values for γ were chosen from the previous experiment.

The ERF is calculated once at the end of each episode and it is defined as the euclidean distance between the hexapod

at time T and its initial position $(0, 0)$ projected onto the xy -plane:

$$R_{\text{ex}} := \sqrt{x_T^2 + y_T^2}, \quad (10)$$

where (x_T, y_T) are the coordinates of the center of the robot in world coordinates at time $t = T$.

The IRF is calculated differently compared to the previous experiment. In a high-dimensional system as the hexapod, it is

not possible to compute the PI of the entire system with a reasonable effort, as the computational cost of $I(S_t; S_{t+1})$ grows exponentially for every new sensor. It would be natural to reduce the computational cost by calculating the PI based on a model of the morphology, but this would violate our claim that the PI incorporates the morphology without the need of explicitly modeling it. Hence, we decided to use the following method to approximate the PI and the entropy H (see **Figure 4D**). Let $S_i(t)$, $i = 1, 2, \dots, 12$, be the angular position sensors for the 12 actuators. We then chose two sensors k, l with $1 \leq k, l \leq 12$, $k \neq l$, randomly from the 12 possible sensors, and calculated

$$\begin{aligned} PI_u &:= I(S_k(t+1), S_l(t+1); S_k(t), S_l(t)) \\ H_u &:= H(S_k(t), S_l(t)). \end{aligned} \quad (11)$$

The overall PI and entropy are then calculated as the sum of n randomly chosen PI_u and H_u pairings, with the additional constraint that each sensor pair k, l appears only once in the approximations. The resulting IRFs are then given by:

$$R_{in}^{PI} := \sum_{u=1}^n PI_u \quad \text{and} \quad R_{in}^H := \sum_{u=1}^n H_u, \quad (12)$$

where n is the number of pairings. For $n > 20$ no difference was found for the approximated PI, which is why $n = 20$ was chosen for the remainder of this work.

The overall reward functions are then given by:

$$R^{\text{PI}} := R_{\text{ex}} + \beta(\gamma)R_{in}^{PI}R^H := R_{\text{ex}} + \beta(\gamma)R_{in}^H \quad (13)$$

where $\beta(\gamma)$ is defined as in the cart-pole swing-up experiment (see Equation 9).

A common recurrent neural network central pattern generator layout is chosen, which can also be found in literature (e.g., Campos et al., 2010; von Twickel et al., 2011; Markelić and Zahedi, 2007), thereby using the same neuron model as in the cart-pole experiment (see above). As all legs in the hexapod are morphologically equivalent, only the synaptic weights of one leg controller are open to parameter adaptation in the PGPE algorithm. The values are then copied to the other leg controllers. This reduces the number of parameters for the entire controller to 32 (see **Figures 4B,C**).

The results (see **Figure 5**) show that neither the PI nor the entropy have a noticeable effect on the learning performance. The mean values of the 100 experiments for each parameter as well as the standard deviation are almost identical. This point will be addressed in the discussion of this work (see Section 4).

3.3. HEXAPOD SELF-RESCUE

The third experiment is designed to combine and extend the two previous experiments. It combines them as a high-dimensional morphology, similar to that used in the locomotion experiment, is trained with an unsteady ERF, which is similar to that used in the cart-pole experiment. It extends the previous experiments as the number of parameters in the controller is a magnitude larger and because an unconventional control structure is used for the desired task. The most distinctive difference to the previous experiments is the non-trivial environment. The next paragraphs will explain the experimental set-up in detail before the section closes with a discussion of the results.

We used the simulated hexapod robot of the LPZROBOTS simulator (Martius et al., 2012). The hexapod has 12 active and 16 passive degrees of freedom (see **Figure 6**). The active joints take the desired next angular position as their input and deliver the current actual angular position as their output. The controller is a fully connected one-layer feed-forward neural network without lateral connections and the hyperbolic transfer function $a_{t+1} = \tanh(Ws_t + v)$, where a_{t+1} and s_t are the next action and the current sensor values, W is the connection matrix, and v is the vector of biases. The resulting controller is parameterized by 156 parameters, 144 for the synaptic weights and 12 for the bias values. Note, that the controller is generic and has no a priori structuring or other robot-specific details.

The task for the hexapod is to rescue itself from a trap. For this purpose, it is placed in a closed rectangular arena (see **Figure 7**). The difficulty of the task is determined by the height of the arena's walls, denoted by $h \in \{0.0\text{m}, 0.1\text{m}, 0.2\text{m}\}$ (see **Figure 6**). For comparison, the length of the lower leg (up to the knees) is 0.45 m. The size-proportion of the robot and the trap can be seen in **Figure 6B**.

The ERF is given by

$$R_{\text{ex}} := \begin{cases} \sqrt{x_T^2 + y_T^2} - r & \text{if } \sqrt{x_T^2 + y_T^2} - r > 0 \\ 0 & \text{otherwise,} \end{cases} \quad (14)$$

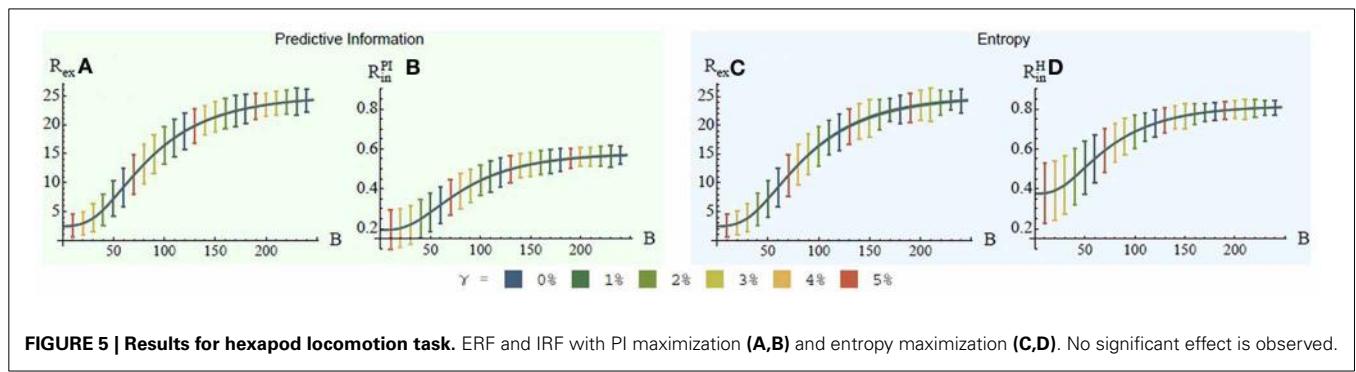
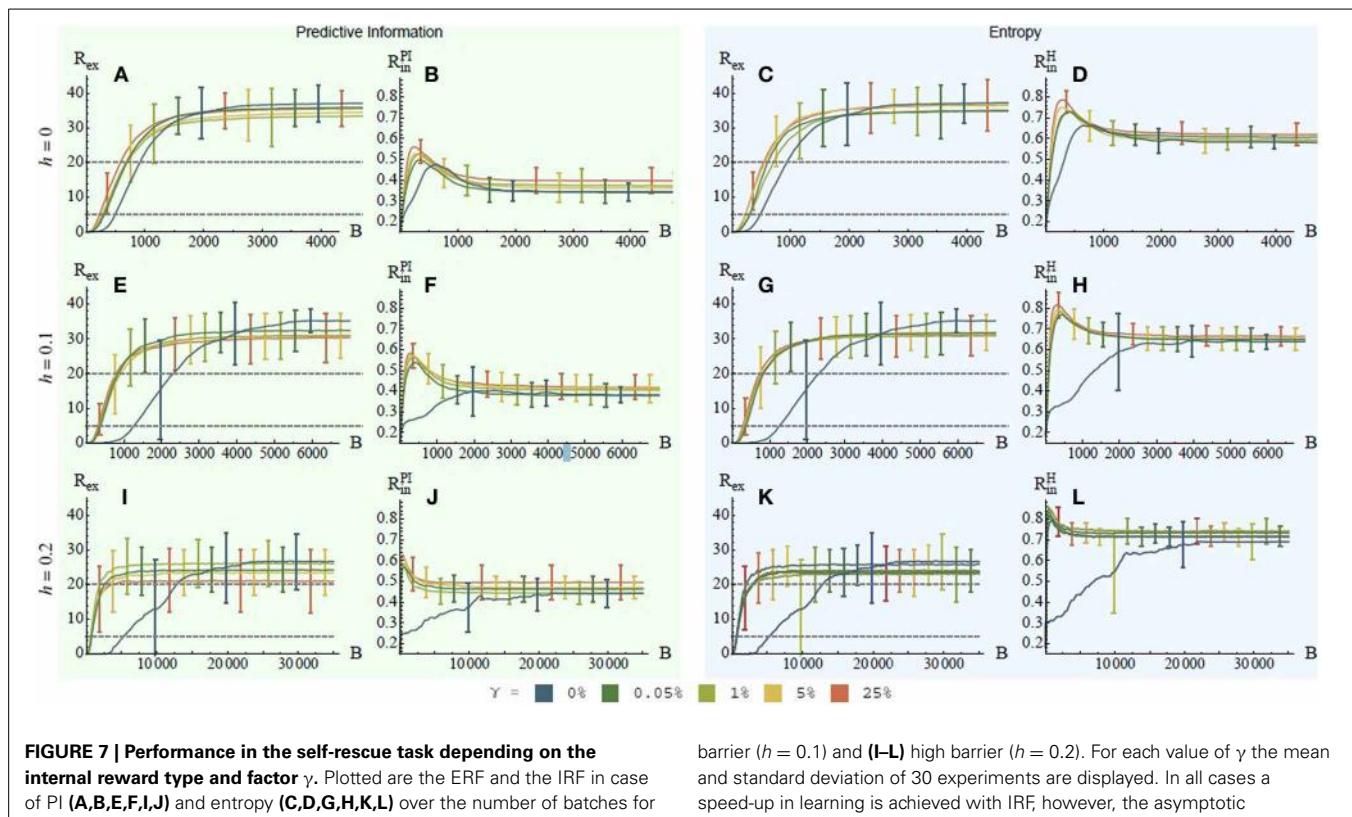
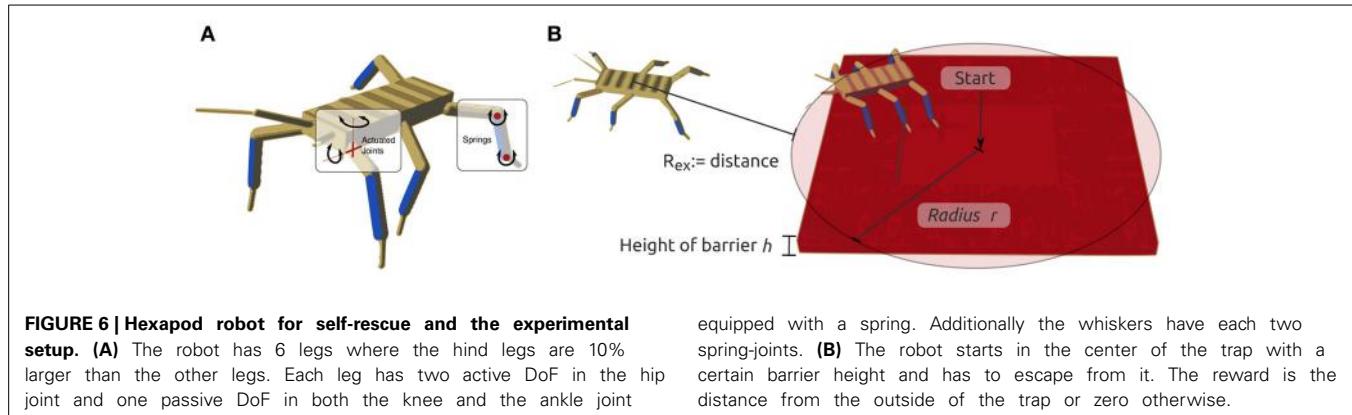


FIGURE 5 | Results for hexapod locomotion task. ERF and IRF with PI maximization (**A,B**) and entropy maximization (**C,D**). No significant effect is observed.



where r is the radius of the trap (Figure 6) and (x_T, y_T) is the position of the center of the robot in world coordinates at the end of a roll-out ($t = T$). The IRFs and overall reward functions are identical to those used in the previous experiment (see Equations (11) and (12)).

As before, the performance of PGPE with $\gamma = 0$ for different values for σ_{init} and α were evaluated, and the best are chosen for presentation here, which are $\sigma_{\text{init}} = 2$ and $\alpha = 0.5$. A different learning rate $\alpha_\sigma = 0.05$ was chosen for the update of σ (see Equation 3). Each episode consisted of $T = 1250$ iterations (25s) with one roll-out per episode. A total of $B = 5000, 7000$, and 35000 batches were conducted for the different heights h .

We compare the performance for different values of the IRF factor $\gamma \in \{0, 0.05, 1, 5, \text{ and } 25\%\}$ and performed 30 experiments for each setting. Figure 7 displays the results. As for the cart-pole experiment, the plots for the PI and entropy in Figure 7 report a clear picture of an exploration phase (high value) followed by an exploitation phase (lower value).

To compare the results, we set two threshold values at $R_{\text{ex}} = 5$ and $R_{\text{ex}} = 20$ which refer to a 5m and 20m distance between the hexapod and the walls of the arena. The first threshold reflects a successful learning of the task, because it means that hexapod reliably escapes the arena. The second threshold represents the case when in addition also a high locomotion speed is achieved

after a successful escape. For the simplicity of argumentation, we compare two cases, i.e., $\gamma = 0\%$ and $\gamma = 1\%$. If there is no wall ($h = 0\text{m}$) the system with IRF $\gamma = 1\%$ requires only half the amount of batches compared to no IRF (250 batches vs. 500 batches, see **Figures 7A,C**). For the arena with a medium height ($h = 0.1\text{m}$), the learning success speed ratio increases to approximately three (350 batches vs. 1100 batches, see **Figures 7E,F**). The results are decisive for the arena with high walls ($h = 0.2\text{m}$), as the system with IRF requires about 1000 batches on average compared to the 5000 batches on average that are required by the systems without IRF (see **Figures 7I,K**).

This leads to the conclusion that both, PI and entropy, are beneficial if the short-term learning success is of the primary interest. However, the asymptotic learning success of those hexapods with IRF is either equal or lower compared to those without an IRF in all experiments. This is valid for the one-step PI and for the entropy. Thus, both are not necessarily beneficial if the long-term, asymptotic learning performance in an episodic policy gradient setting is important.

4. DISCUSSION

This paper discussed the one-step PI (Bialek et al., 2001) as an information-driven intrinsic reward in the context of an episodic policy gradient method. The reward is considered to be intrinsic, because it is task-independent and it relies only on the information of the sensors of an agent, which, by definition, represent the agent's intrinsic view on the world. We chose the maximization of the one-step PI as an IRF, because it has proved to encourage behaviors which show properties of morphological computation without the need to model the morphology (Zahedi et al., 2010).

The IRF was linearly combined with a task-dependent ERF in an episodic RL setting. Specifically, PGPE (Sehnke et al., 2010) was chosen as RL method, because it allows to learn arbitrary policy parametrizations. Within this set-up, three different types of experiments were performed. The following paragraph will summarize the results before they are discussed.

The first experiment was the learning of the cart-pole swing-up task. Four controllers were evaluated of which three were less successful and one showed good results. The ERF was designed to be difficult to maximize without the IRF, and the task contradicted the maximization of the entropy and PI. The best controller did not show a significant improvement of the learning performance with respect to its asymptotic behavior. An improvement could only be observed during the first learning steps. Moreover, the choice of the linear combination factor γ is critical. For all controllers a minor and not significant improvement is observable. In case of the entropy maximization, any factor $\gamma > 0\%$ showed an improvement in learning speed and learning performance.

A locomotion behavior was learned for a hexapod in the second experiment. The entire set-up used well-known components for the environment, modular controller, ERF, and morphology so that the task was solved without IRF in only a few learning steps. No effect of the PI and entropy was observed.

The third experiment combined the previous two and extended them by a non-trivial environment. A hexapod had to escape from a trap and was only rewarded outside of it. The

results showed no significant difference between the PI and the entropy as IRFs. The learning speed was significantly improved by both IRFs with increasing difficulty of the task. The asymptotic performance was either equal or worse when an IRF was introduced.

The hexapod locomotion experiment teaches us that the information-theoretic reward functions (PI and entropy) has no effect in well-defined experimental set-ups.

The cart-pole and the hexapod self-rescue experiments teach us that the maximal values of the IRF should be around one percent of the maximal ERF value to improve the learning speed and learning performance in the short-term. The asymptotic behavior is either not or negatively effected by the one-step PI. The cart-pole experiment indicates that maximizing the entropy is superior to maximizing the PI, whereas the hexapod self-rescue does not show such a clear picture. The success of the entropy in both experiments is explained by the ERFs. Due to their nature, random changes in the policy parameters are unlikely to result in changes in the ERF during the first batches. Hence, maximizing the entropy results in an exploration until the ERF is triggered.

The PI, defined as the entropy over the sensor states subtracted by the conditional entropy of consecutive sensor states does not result in superior results for the cart-pole compared to just using the entropy for the following reason. In this set-up, the morphology and environment are very simple and deterministic, and therefore, do not produce any noise or other uncertainties in the sensor data stream. The uncertainty about the next possible angular position of the pole is small, if the current pole position is known. In other words, the cart-pole system is regular by definition and no further regularities can be found by maximizing the PI. We speculate that the conditional entropy, which cannot be reduced by the learning in this setting, dampens the exploration effect of the entropy term in the PI maximization. For the hexapod rescue experiment, the situation is different. There is an uncertainty about the next sensor state, given the current sensor state which result from the morphology and the construction of the arena. The PI maximization is able to find regularities which can then be exploited to maximize the ERF in the RL setting.

The results contradict our intuition, as the one-step predictive information has shown good results when applied as an information-driven self-organization principle in the context of embodied artificial intelligence (Zahedi et al., 2010; Martius et al., 2013). The intuitively plausible next step was to guide the information-driven self-organization toward solving a goal by combining it with an external reward signal in a reinforcement learning context. The approach evaluated in this paper was to linearly combine the PI with an external reward signal in an episodic policy gradient learning. If anything, then the PI showed positive short-term results, if the world was considerably probabilistic and if the external reward was sparse. Compared to no intrinsic reward the PI showed negative results for its asymptotic behavior. The performance of the PI was either equal or worse compared to the entropy in all cases. This leads to the conclusion that research in the context of information-driven intrinsic rewards and reinforcement learning should be carried out in other directions, which are briefly described in the final paragraph.

We have used a constant combination factor γ for all experiments presented in this work. It is known from general learning theory that a decaying learning rate is required for the convergence of a system. We chose not to use a decaying learning factor, because this means that the internal drive is slowly damped until its effect is neglectable (at least in a technical application). This would contradict the idea of motivation-driven and open-ended learning of embodied agents. However, the results of our present paper reveal a disadvantage of this approach in the asymptotic limit, and therefore, suggest, contrary to our original thoughts, to pursue a strategy with a decaying combination factor. The second possible modification of this approach is to exchange the linear combination of the internal and external reward by a non-linear function, of which multiplicative and exponential functions are two examples. Third, using a gradient of the PI instead of a random exploration in the context

of RL is a promising approach that is currently investigated. In this approach, we will use a gradient on an estimate of the PI and not the error of a predictor as in e.g., (Schmidhuber, 1991). Fourth, we will continue to evaluate other information-theoretic measures in the context of task-dependent learning with the support of information-driven intrinsic motivation. In addition to using correlation measures, such as the mutual information, we believe that causal measures in the sensorimotor loop (Ay and Zahedi, 2013), such as the measure considered in (Zahedi and Ay, 2013), are good candidates for future research in this field.

ACKNOWLEDGMENTS

This work was funded by the German Priority Program *Autonomous Learning* (DFG-SPP 1527). We would like to thank the reviewers for their very helpful comments.

REFERENCES

- Ay, N., and Polani, D. (2008). Information flows in causal networks. *Adv. Complex Syst.* 11, 17–41. doi: 10.1142/S0219525908001465
- Ay, N., and Zahedi, K. (2013). “An information theoretic approach to intention and deliberative decision-making of embodied systems,” in *Advances in Cognitive neurodynamics III* ed Y. Yamaguchi (Heidelberg: Springer).
- Ay, N., Bertschinger, N., Der, R., Gütter, F., and Olbrich, E. (2008). Predictive information and explorative behavior of autonomous robots. *Eur. Phys. J. B* 63, 329–339. doi: 10.1140/epjb/e2008-00175-0
- Barto, A. G., Singh, S., and Chentanez, N. (2004). “Intrinsically motivated learning of hierarchical collections of skills,” in *Proceedings of 3rd International Conference on Developmental Learning*, (San Diego, CA), 112–119.
- Barto, A. G., Sutton, R. S., and Anderson, C. W. (1983). Neuron-like adaptive elements that can solve difficult learning control problems. *IEEE Trans. Syst. Man. Cybern. SMC-13*, 834–846. doi: 10.1109/TSMC.1983.6313077
- Bellman, R. E. (2003). *Dynamic Programming*. Mineola, NY: Dover Publications, Incorporated.
- Bialek, W., Nemenman, I., and Tishby, N. (2001). Predictability, complexity, and learning. *Neural Comput.* 13, 2409–2463. doi: 10.1162/08997601753195969
- Brooks, R. A. (1991). “Intelligence without reason,” in *Proceedings of the 12th International Joint Conference on Artificial Intelligence (IJCAI-91)*, eds J. Myopoulos, and R. Reiter, (San Mateo, Sydney: Morgan Kaufmann publishers Inc.), 569–595.
- Campos, R., Matos, V., and Santos, C. (2010). “Hexapod locomotion: a nonlinear dynamical systems approach,” in *Conference Of IEEE Industrial Electronics. Proceedings*, (Glendale, AZ), 1546–1551.
- Clark, A. (1996). *Being There: Putting Brain, Body, and World Together Again*. Cambridge, MA: MIT Press.
- Cover, T. M., and Thomas, J. A. (2006). *Elements of Information Theory*, Vol. 2. Hoboken, NJ: Wiley.
- Crutchfield, J. P., and Young, K. (1989). Inferring statistical complexity. *Phys. Rev. Lett.* 63, 105–108. doi: 10.1103/PhysRevLett.63.105
- Cuccu, G., Luciw, M., Schmidhuber, J., and Gomez, F. (2011). “Intrinsically motivated evolutionary search for vision-based reinforcement learning,” in *Proceedings of the 2011 IEEE Conference on Development and Learning and Epigenetic Robotics IEEE-ICDL-EPIROB*, (Frankfurt: IEEE).
- Dayan, P., and Balleine, B. W. (2002). Reward, motivation, and reinforcement learning. *Neuron* 36, 285–298. doi: 10.1016/S0896-6273(02)00963-7
- Der, R., and Martius, G. (2012). *The Playful Machine: Theoretical Foundation and Practical Realization of Self-Organizing Robots (Cognitive Systems Monographs)*. Berlin; Heidelberg: Springer.
- Doya, K. (2000). Reinforcement learning in continuous time and space. *Neural Comput.* 12, 219–245. doi: 10.1162/089976000300015961
- Geva, S., and Sitte, J. (1993). The cart pole experiment as a benchmark for trainable controllers. *IEEE Control Syst. Mag.* 13, 40–51. doi: 10.1109/37.236324
- Grassberger, P. (1986). Toward a quantitative theory of self-generated complexity. *Int. J. Theor. Phys.* 25, 907–938. doi: 10.1007/BF00668821
- Kaplan, F., and Oudeyer, P.-Y. (2004). “Maximizing learning progress: an internal reward system for development,” in *Embodied Artificial Intelligence*, eds F. Iida, R. Pfeifer, L. Steels, and Y. Kuniyoshi (Berlin; Heidelberg: Springer-Verlag), 259–270.
- Klyubin, A. S., Polani, D., and Nehaniv, C. L. (2004). “Organization of the information flow in the perception-action loop of evolved agents,” in *Proceedings of the 2004 NASA/DoD Conference on Evolvable Hardware, 2004*, (Seattle, WA), 177–180.
- Little, D. Y., and Sommer, F. T. (2013). Learning and exploration in action-perception loops. *Front. Neural Circuits* 7:37. doi: 10.3389/fncir.2013.00037
- Markelić, I., and Zahedi, K. (2007). “An evolved neural network for fast quadrupedal locomotion,” in *Advances in Climbing and Walking Robots, Proceedings of 10th International Conference (CLAWAR 2007)*, eds M. Xie and S. Dubowsky, (World Scientific Publishing Company), 65–72.
- Martius, G., Der, R., and Ay, N. (2013). Information driven self-organization of complex robotic behaviors. *PLoS ONE* Singapore 8:e63400. doi: 10.1371/journal.pone.0063400
- Martius, G., and Herrmann, J. M. (2012). Variants of guided self-organization for robot control. *Theory Biosci.* 131, 129–137. doi: 10.1007/s12064-011-0141-0
- Martius, G., Hesse, F., Gütter, F., and Der, R. (2012). LPZROBOTS: a free and powerful robot simulator, version 0.7. Available online at: <http://robot.informatik.uni-leipzig.de/software>
- Oudeyer, P.-Y., Kaplan, F., and Hafner, V. V. (2007). Intrinsic motivation systems for autonomous mental development. *IEEE Trans. Evol. Comput.* 11, 265–286. doi: 10.1109/TEVC.2006.890271
- Pasemann, F., Steinmetz, U., and Dieckman, U. (1999). “Evolving structure and function of neurocontrollers,” in *Proceedings of the Congress Evolutionary Computation CEC 99*, Vol. 3, (Washington, DC).
- Pfeifer, R., and Bongard, J. C. (2006). *How the Body Shapes the Way We Think: A New View of Intelligence*, (Cambridge, MA: The MIT Press; Bradford Books).
- Pfeifer, R., Lungarella, M., and Iida, F. (2007). Self-organization, embodiment, and biologically inspired robotics. *Science* 318, 1088–1093. doi: 10.1126/science.1145803
- Prokopenko, M., Gerasimov, V., and Tanev, I. (2006). “Evolving spatiotemporal coordination in a modular robotic system,” in *Proceedings on SAB’06*, Vol. 4095, (Rome, Italy), 558–569.
- Schmidhuber, J. (1990). “A possibility for implementing curiosity and boredom in model-building neural controllers,” in *Proceedings of SAB’90*, (Cambridge, MA), 222–227.
- Schmidhuber, J. (1991). “Curious model-building control systems,” in *In Proceedings on International Joint Conference on Neural Networks, Singapore*, (IEEE), 1458–1463.
- Schmidhuber, J. (2006). Developmental robotics, optimal artificial curiosity, creativity, music, and the fine arts. *Connect. Sci.* 18, 173–187. doi: 10.1080/09540090600768658

- Sehnke, F., Osendorfer, C., Rückstiess, T., Graves, A., Peters, J., and Schmidhuber, J. (2010). Parameter-exploring policy gradients. *Neural Netw.* 23, 551–559. doi: 10.1016/j.neunet.2009.12.004
- Storck, J., Hochreiter, S., and Schmidhuber, J. (1995). “Reinforcement driven information acquisition in non-deterministic environments,” in *Proceedings of the International Conference on Artificial Neural Networks*, Vol. 2, (Paris: EC2 & Cie), 159–164.
- Sutton, R. S., and Barto, A. G. (1998). *Reinforcement Learning: An Introduction*, Cambridge, MA: MIT Press.
- von Twickel, A., Büschges, A., and Pasemann, F. (2011). Deriving neural network controllers from neuro-biological data: implementation of a single-leg stick insect controller. *Biol. Cybern.* 104, 95–119. doi: 10.1007/s00422-011-0422-1
- von Uexküll, J. (1934). “A stroll through the worlds of animals and men,” in *Instinctive Behavior*, ed C. H. Schiller, (New York, NY: International Universities Press), 5–80.
- Yi, S., Gomez, F., and Schmidhuber, J. (2011). “Planning to be surprised: optimal Bayesian exploration in dynamic environments,” in *Proceedings on Fourth Conference on Artificial General Intelligence (AGI)*, (Mountain View, CA: Google).
- Zahedi, K., and Ay, N. (2013). Quantifying morphological computation. *Entropy* 15, 1887–1915. doi: 10.3390/e15051887
- Zahedi, K., Ay, N., and Der, R. (2010). Higher coordination with less control—a result of information maximization in the sensori-motor loop. *Adapt. Behav.* 18, 338–355. doi: 10.1177/1059712310375314
- Zahedi, K., von Twickel, A., and Pasemann, F. (2008). “Yars: a physical 3d simulator for evolving controllers for real robots,” in *SIMPAR 2008* Vol. 5325, eds S. Carpin, I. Noda, E. Pagello, M. Reggiani, and O. von Stryk (Berlin; Heidelberg: Springer), 71–82.
- Citation:** Zahedi K, Martius G and Ay N (2013) Linear combination of one-step predictive information with an external reward in an episodic policy gradient setting: a critical analysis. *Front. Psychol.* 4:801. doi: 10.3389/fpsyg.2013.00801
- This article was submitted to Cognitive Science, a section of the journal *Frontiers in Psychology*.
- Copyright © 2013 Zahedi, Martius and Ay. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 21 June 2013; *accepted:* 10 October 2013; *published online:* 04 November 2013.



Incremental learning of skill collections based on intrinsic motivation

Jan H. Metzen^{1*} and Frank Kirchner^{1,2}

¹ Robotics Research Group, Faculty 3 – Mathematics and Computer Science, Universität Bremen, Bremen, Germany

² Robotics Innovation Center, German Research Center for Artificial Intelligence (DFKI), Bremen, Germany

Edited by:

Tom Stafford, University of Sheffield, UK

Reviewed by:

Antonio Novellino, ett s.r.l., Italy

Frank Van Der Velde, University of Twente, Netherlands

***Correspondence:**

Jan H. Metzen, Robotics Research Group, Faculty 3 - Mathematics and Computer Science, Universität Bremen, Robert-Hooke-Str. 5, Bremen, 28359, Germany
e-mail: jhm@informatik.uni-bremen.de

Life-long learning of reusable, versatile skills is a key prerequisite for embodied agents that act in a complex, dynamic environment and are faced with different tasks over their lifetime. We address the question of how an agent can learn useful skills efficiently during a developmental period, i.e., when no task is imposed on him and no external reward signal is provided. Learning of skills in a developmental period needs to be incremental and self-motivated. We propose a new incremental, task-independent skill discovery approach that is suited for continuous domains. Furthermore, the agent learns specific skills based on intrinsic motivation mechanisms that determine on which skills learning is focused at a given point in time. We evaluate the approach in a reinforcement learning setup in two continuous domains with complex dynamics. We show that an intrinsically motivated, skill learning agent outperforms an agent which learns task solutions from scratch. Furthermore, we compare different intrinsic motivation mechanisms and how efficiently they make use of the agent's developmental period.

Keywords: hierarchical reinforcement learning, skill discovery, intrinsic motivation, life-long learning, graph-based representation

1. INTRODUCTION

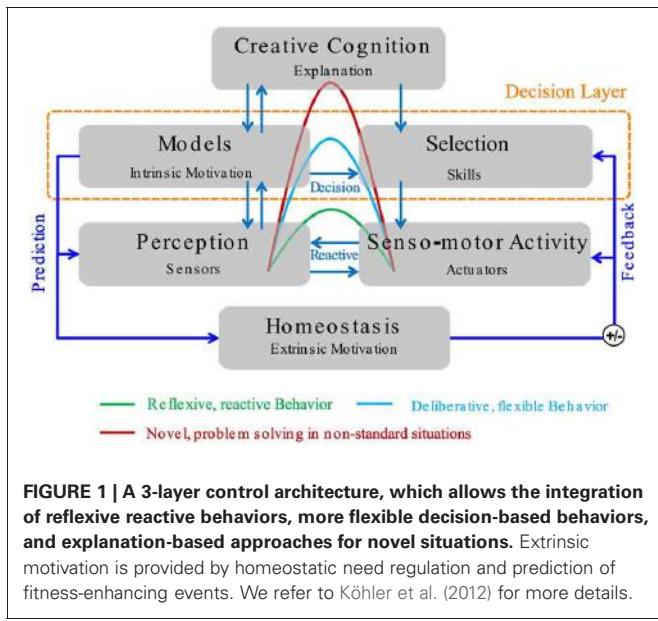
Embodied agents like robots are used in increasingly complex, real-world domains, such as domestic and extraterrestrial settings. A simple, reactive control approach is not sufficient as it lacks the ability to predict and control the environment on larger scales of time and space. For this, agents must be able to build up competencies and knowledge about the world and store these in a convenient way so that they can be accessed fast and reliably. This requires control architectures which allow, *inter alia*, model-learning, predictive control, learning reusable skills, and even the integration of high-level cognitive elements. See **Figure 1** for an example of such an architecture.

In this work, we focus on the middle, “decision” layer of such an architecture. One main objective on this layer is to learn a *repertoire of reusable skills*. Such skills may be the ability to reliably grasp objects, to throw, catch, or hit a ball, or to use a tool for a specific task like using a hammer to drive a nail into a wall. A repertoire of skills is useful for embodied agents which have to solve several different but related tasks during their lifetime. Instead of learning every novel task from scratch, learning skills allows that acquired capabilities are reused, i.e., transferred between tasks. Furthermore, being able to use prelearned skills may dramatically increase response times and therefore reduce the probability of system failure. One approach to skill learning is hierarchical reinforcement learning (Barto and Mahadevan, 2003), which has been applied successfully in robotic applications (see, e.g., Kirchner, 1998). Since the acquired skills shall be reusable, they should not be driven by external, task-specific reward. Instead, the agent should learn skills in a task-independent manner. In addition, an autonomous agent

must decide on its own what constitutes a useful skill; this is denoted as *skill discovery*.

Existing skill discovery approaches are mostly tailored to discrete domains or to decomposing a specific task into sub-tasks. While the former have limited significance for continuous domains like robotics, the latter might yield skills that are task-specific and not reusable. The main contribution of this paper is a new skill discovery method which is suited for continuous domains and does not require external tasks and rewards. This method allows the agent to generate a collection of skills during a *developmental period*, in which the agent can explore freely without having to maximize external reward. The proposed skill discovery method is based on an incremental, hierarchical clustering of a learned state transition graph. This graph encodes the structure and dynamics of a domain. Densely connected subgraphs (“clusters”) of this graph correspond to qualitatively similar situations in the domain. Skills are learned for transitioning from one cluster to an adjacent one, i.e., for purposefully reaching a specific configuration of the domain.

In large domains with complex dynamics, exploring the environment, which is a prerequisite for skill discovery, is challenging by itself as is the decision whether the agent should engage in skill learning or exploration. We consider *intrinsic motivation* to reward the agent for (a) exploring novel parts of the environment and for (b) engaging in learning skills whose predictive model exhibits large error. We define novelty with regard to a set of observed states and predict skill effects based on a learned skill model which allows predicting state transitions conditioned on the specific skill.



We present an empirical analysis of the proposed approach in two continuous, high-dimensional domains with complex dynamics. We evaluate empirically to which extent the agent can benefit from reusing skills, which influence the specific skill discovery approach and the definition of intrinsic motivation have onto the agent's performance, and how the length of the agent's developmental period affects the task performance. Furthermore, we present evidence that the intrinsic motivation mechanisms can identify how much time should be spent on learning specific skills.

The paper is structured as follows: section 2 provides the necessary background and summarizes some of the most closely related works. Section 3 gives details of the main methodological contributions of this paper. In section 4, we present and discuss the results obtained in the empirical analysis. In section 5, we draw a conclusion and provide an outlook.

2. BACKGROUND AND RELATED WORK

In this section, we present briefly the required background in hierarchical reinforcement learning and give a review of related works in the areas of skill discovery and intrinsic motivation.

2.1. HIERARCHICAL REINFORCEMENT LEARNING

Computational Reinforcement Learning (RL) (Sutton and Barto, 1998) refers to a class of learning methods that aims at learning behavior policies which are optimal with regard to a reward signal, through interaction with an environment. The most popular problem class for RL are Markov Decision Processes (MDPs). An MDP M can be formalized as a 4-tuple $M = (S, A, P_{ss'}^a, R_{ss'}^a)$ where S is a set of states of the environment, A is a set of actions, $P_{ss'}^a = P(s_{t+1} = s' | s_t = s, a_t = a)$ is the 1-step state transition probability also referred to as the "dynamics," and $R_{ss'}^a = E\{r_{t+1} | s_t = s, a_t = a, s_{t+1} = s'\}$ is the expected reward. In RL, these quantities are usually unknown to the agent but can be estimated based on samples collected during exploration. If both S

and A are finite, we call M a discrete MDP, otherwise we call it a continuous MDP. The goal of RL is to learn without explicit knowledge of M a policy π^* such that some measure of the long-term reward is maximized. Popular approaches to RL include value-function based methods, which are based on approximating the optimal action-value function $Q^*(s, a) = \sum_{s'} P_{ss'}^a [R_{ss'}^a + \gamma \max_{a'} Q^*(s', a')]$, where $\gamma \in [0, 1]$ is a discount factor, and direct policy search methods, which search directly in the space of policies based on, e.g., evolutionary computation (Whiteson, 2012).

This paper focuses on learning *skills* using Hierarchical RL (Barto and Mahadevan, 2003). In Hierarchical RL, behavior is not represented by a monolithic policy but by a hierarchy of policies, where policies on the lowest layer correspond to simple skills and policies on higher layer are based on these skills and represent more complex behavior. One popular approach to Hierarchical RL is the *options framework* (Sutton et al., 1999). An option o is the formalization of a temporally extended action or skill and consists of three components: the option's initiation set $I_o \subset S$ that defines the states in which the option may be invoked, the option's termination condition $\beta_o : S \rightarrow [0, 1]$ which specifies the probability of option execution terminating in a given state, and the option's policy π_o which defines the probability of executing an action in a state under option o . In the options framework, a policy on a higher layer may in any state s decide not solely to execute a primitive action but also to call any of the lower-layer options for which $s \in I_o$. If an option is invoked, the option's policy π_o is followed for several time steps until the option terminates according to β_o . The option's policy π_o is defined relative to an option-specific "pseudo-reward" function R_o that rewards the option for achieving the skill's objective. Skill learning denotes learning π_o given I_o , β_o , and R_o . Skill discovery, on the other hand, requires choosing appropriate I_o , β_o , and R_o for a new option o . Skill discovery is very desirable since the quantities I_o , β_o , and R_o need not be predefined but can be identified by the agent itself and thus, skill discovery increase the agent's autonomy. We give a review of related works in the next section.

2.2. SKILL DISCOVERY

Most prior work on autonomous skill discovery is based on the concept of *bottleneck areas* in the state space. Informally, bottleneck areas have been described as the border states of densely connected areas in the state space (Menache et al., 2002) or as states that allow transitions to a different part of the environment (Şimşek and Barto, 2004). A more formal definition is given by Şimşek and Barto (2009), in which bottleneck areas are states that are local maxima of betweenness—a measure of centrality on graphs—on the transition graph. Once bottleneck areas have been identified, typically one (or several) skills are defined that try to reach this bottleneck, i.e., that terminate with positive pseudo-reward if the bottleneck area is reached, can be invoked in a local neighborhood of the bottleneck, and terminate with a negative pseudo-reward when departing too far from the bottleneck.

Since betweenness requires complete knowledge of the transition graph and is computationally expensive, several heuristics have been proposed to identify bottlenecks. One class of heuristics are *frequency-based approaches* that compute local statistics of

states like diverse density (McGovern and Barto, 2001) and relative novelty (Şimşek and Barto, 2004). An other class of heuristics that is typically more sample-efficient are *graph-based approaches* which are based on estimates of the domain's state transition graph. Graph-based approaches to skill discovery aim at partitioning this graph into subgraphs which are densely connected internally but only weakly connected with each other. Menache et al. (2002) propose a top-down approach for partitioning the global transition graph based on the max-flow/min-cut heuristic. Şimşek et al. (2005) follow a similar approach but partition local estimates of the global transition graph using a spectral clustering algorithm and use repeated sampling for identifying globally consistent bottlenecks. Mannor et al. (2004) propose a bottom-up approach that partitions the global transition graph using agglomerative hierarchical clustering. Metzen (2012) proposes an extension of this approach called OGAHC. OGAHC is incremental and can thus be performed several times during the learning process. A further approach for identifying bottlenecks is to monitor the propagation of Q-values in the planning phase of a model-based RL architecture. For instance, Kirchner and Richter (2000) have shown that the so-called significance values become large close to bottlenecks of the domain.

Relatively few works on autonomous skill discovery in domains with continuous state spaces exist. Frequency-based approaches do not easily generalize to such domains since their statistics are typically related to individual states and there exist infinitely many such states in continuous domains. Similarly, the 1-to-1 relationship between states and graph nodes hinders the direct applicability of graph-based approaches. Mannor et al. (2004) have evaluated their agglomerative hierarchical clustering approach in the mountain car domain by uniformly discretizing the state space. However, this uniform discretization is suboptimal since it suffers from alignment effects and the "curse of dimensionality." Learning an adaptive discretization in the form of a transition graph that captures the domain's dynamics using the FIGE heuristic (see section 3.2) is shown to perform considerably better (Metzen, in press). However, FIGE is a batch method and requires that skill discovery is performed at a prespecified point in time.

One skill discovery method that has been designed for continuous domains is "skill chaining" (Konidaris and Barto, 2009). Skill chaining produces chains (or more general: trees) of skills such that each skill allows reaching a specific region of the state space, such as a terminal region or a region where an other skill can be invoked. In which region of the state space a skill can be invoked depends mainly on the representability and learnability of the skill in the specific learning system and not directly on concepts like bottlenecks or densely connected regions. Skill chaining requires to specify an area of interest (typically the terminal region of the state space) which is used as target for the skill at the root of the tree. For multi-task domains with several goal regions or domains without a goal region, it is unclear how the root of the skill tree should be chosen.

2.3. LIFELONG LEARNING AND INTRINSIC MOTIVATION

Thrun (1996) suggested the notion of *lifelong learning* in the context of supervised learning for object recognition. In lifelong

learning, a learner experiences a sequence of different but related tasks. Due to this relatedness, learned knowledge can be transferred across multiple learning tasks, which can allow generalizing more accurately from less training data. The concept of lifelong learning was extended to RL by, e.g., Sutton et al. (2007). In RL, lifelong learning is often combined with *shaping*, which denotes a process where a trainer rewards an agent for a behavior that progresses toward a desired target behavior which solves a complex task. Thus, shaping can be seen as a training procedure for guiding the agent's learning process. Shaping was originally proposed in psychology as an experimental procedure for training animals (Skinner, 1938) and has been adopted for training of artificial systems later on (Randløv Alstrøm, 1998). One disadvantage of shaping is that an external teacher is required which selects tasks of a specific complexity carefully by taking the current developmental state of the agent into account. This reduces the agent's autonomy.

A different approach to lifelong learning, in which no external teacher is required, is to provide the agent with a means for *intrinsic motivations*. The term "intrinsically motivated" stems from biology and one of its first appearances was in a paper by Harlow (1950) on the manipulation behavior of rhesus monkeys. According to Baldassarre (2011) "extrinsic motivations guide learning of behaviors that directly increase (evolutionary) fitness" while "intrinsic motivations drive the acquisition of knowledge and skills that contribute to produce behaviors that increase fitness only in a later stage." Thus, similar to shaping, intrinsic motivations contribute to learning not as a learning mechanism *per se*, but rather as a guiding mechanism which guides learning mechanisms to acquire behaviors that increase fitness. According to Baldassarre "(intrinsic motivations) drive organisms to continue to engage in a certain activity if their competence in achieving some interesting outcomes is improving, or if their capacity to predict, abstract, or recognize percepts is not yet good or is improving...." Accordingly, learning signals produced by intrinsic motivations tend to decrease or disappear once the corresponding skill is acquired.

Computational approaches to intrinsic motivation [see Oudeyer and Kaplan (2007) for a typology] have become popular in hierarchical RL in the last decade resulting in the area of Intrinsically Motivated Reinforcement Learning (IMRL) (Barto et al., 2004). Work on intrinsic motivation in RL, however, dates back to the early 1990s (Schmidhuber, 1991). IMRL often employs a *developmental setting* [see, e.g., Stout and Barto (2010) and Schembri et al. (2007)], which differs slightly from the usual RL setting where the objective is to maximize the accumulated external reward. In the developmental setting, the agent is given a developmental period, which can be considered as its "childhood," in which no external reward is given to the agent. This allows the agent to explore its environment freely without having to maximize the accumulated reward (exploitation). On the other hand, the agent is not guided by external reward but needs to have a means for intrinsic motivation. The objective in the developmental setting is to learn skills which allow to quickly learn high-quality policies in tasks that are later imposed onto the agent. Thus, the objective can be seen as a kind of optimal exploration for skill learning, in contrast to finding the optimal balance

between exploration and exploitation as in usual RL. Different mechanisms for intrinsic motivation have been proposed. A complete review is beyond the scope of this paper, we discuss a selected subset of methods and refer to Oudeyer et al. (2007) for a review.

Barto et al. (2004) investigate how a hierarchically organized collection of reusable skills can be acquired based on intrinsic reward. Their notion of intrinsic reward is based on the novelty response of dopamine neurons. More precisely, the intrinsic reward for a *salient event* is proportional to the error of predicting this salient event based on a learned skill model for this event. This skill model is not only a passive model of the environment but it is also dependent on the agent's action preferences. As a result of the intrinsic reward, once the agent encounters an unpredicted salient event, it is driven to attempt to achieve this event until it has learned to predict it satisfactorily.

Oudeyer et al. (2007) propose an intrinsic motivation system that encourages the robot to explore situations in which its current *learning progress* is maximized. More specifically, the robot obtains a positive intrinsic reward for situations in which the error rate of internal predictive models decreases and a negative one for situations in which it increases. Thereby, the robot focuses on exploring situations whose complexity matches its current stage of development, i.e., situations which are neither too complex (too unpredictable) nor too simple (too predictable).

Hester and Stone (2012) propose a model-based approach for a developing, curious agent called TEXPLORE-VANIR. This approach uses two kinds of intrinsic reward that are derived from the learned model. The first one rewards the agent for exploring parts of the environment for which the variance in the model's prediction is large while the second one rewards the agent for exploring parts of the environment that are *novel* to the agent. The authors show empirically that these intrinsic rewards are helpful for an agent in a developmental setting. Furthermore, the intrinsic rewards also improve the performance of an agent faced with an external task from the very beginning by providing a reasonable explorative bias.

Stout and Barto (2010) propose "competence progress motivation," which generates intrinsic rewards based on the skill competence progress, i.e., how strongly the agent's competence to achieve self-determined goals progresses. The authors show on a simple problem that the approach is able to focus learning efforts onto skills that are neither too simple not too difficult at the moment. While the authors predefine the set of skills that shall be learned, they note that "identifying what skills should be learned is a very important problem and one that a complete motivational system would address." This problem is addressed in this paper.

Note that intrinsic motivations need not be the only source of motivation in a biologically-inspired robotic control architecture such as the one shown in **Figure 1**; rather, extrinsic motivations based on homeostatic need regulation and prediction of fitness-enhancing visceral-body changes (compare Baldassarre, 2011) should be taken into account as well. However, since we focus on the "decision" layer of the architecture, we do not consider these kinds of motivations in detail here.

3. METHODS

In this section, we present an architecture for an IMRL-agent and propose new methods for skill discovery and intrinsic motivation.

3.1. AGENT ARCHITECTURE

We consider an agent situated in an environment with state space S and action space A . We are particularly interested in problems where the state and/or the action space are continuous, more specifically where $S \subseteq \mathbb{R}^{n_s}$ and/or $A \subseteq \mathbb{R}^{n_a}$. We assume that the state transitions (the effects of executing an action in a state) have the Markov property. During its lifetime, the agent may be faced with different tasks in this environment; we assume that each task \mathcal{T}_j is specified by a reward function $\mathcal{R}_j = E(r_{t+1}|s_t = s, a_t = a, s_{t+1} = s')$ and the agent needs to maximize a long-term notion of this reward. Note that each task thus corresponds to a MDP $\mathcal{M}_j = (S, A, \mathcal{P}, \mathcal{R}_j)$, where all tasks share S, A , and \mathcal{P} .

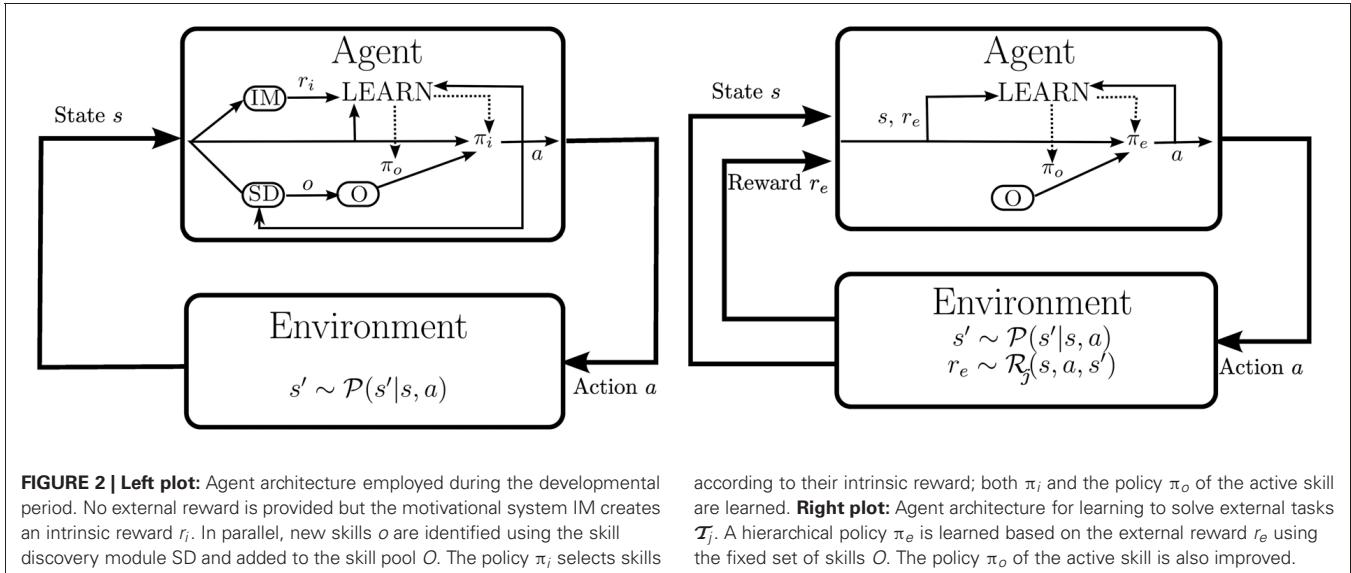
We adopt the developmental setting of IMRL (see section 2.3), i.e., we assume that the agent has a developmental period before it is faced with an external task. The agent-environment interaction during the developmental period can be modeled as an MDP without reward $\mathcal{M}_{\setminus \mathcal{R}} = (S, A, \mathcal{P})$. Thus, we implicitly assume that the developmental period takes place in the same environment where the agent has to solve tasks later on, i.e., we assume S , A , and \mathcal{P} to be identical. While no external objective is imposed on the agent, the agent should use the developmental period nevertheless for learning a repertoire of skills O that can later on help in solving tasks \mathcal{T}_j . Furthermore, we do not provide the agent with a set of subgoals or salient events but require the agent to identify these on its own.

For this, two questions need to be addressed: (a) how are useful and task-independent skills identified autonomously? and (b) how does the agent select actions and skills when no external reward is available? We address these questions in section 3.2 and section 3.3, respectively. For now, we assume that two modules for intrinsic motivation (IM) and skill discovery (SD) exist where IM generates an intrinsic reward signal r_i which the agent uses in place of external reward and SD identifies new skills which are added to the skill repertoire O and whose policy is learned later on by the agent using option learning. The agent's internal architecture during its developmental period is depicted in the left diagram in **Figure 2**.

Once an external task \mathcal{T}_j is imposed onto the agent, the intrinsic reward and the skill discovery modules are disabled, and the agent learns a hierarchical policy π_e over the set of discovered skills O that maximizes the external reward r_e (see right diagram in **Figure 2**). Note that the agent continues to learn option policies π_o based on experience collected; however, the external reward is ignored in skill learning such that options remain task-independent.

3.2. ITERATIVE GRAPH-BASED SKILL DISCOVERY

A skill discovery method which can be used in the outlined architecture needs to exhibit the following properties: (1) it needs to be suited for continuous domains, (2) it needs to be incremental, i.e., the agent must be able to identify new skills at any time and not just once after some predefined amount of



experience was collected, and (3) it must not require that an external reward signal or a goal region of a task exist. None of the methods discussed in section 2.2 fulfills all these requirements. In this work, we propose IFIGE, an incremental extension of FIGE (Metzen, in press), which is combined with an extension of OGAHC (Metzen, 2012) to continuous domains. This combination fulfills all of requirements given above. The key idea of the approach is that a transition graph, which captures the domain's dynamics, is learned incrementally from experience using FIGE and that the learned graphs are clustered into densely connected subgraphs using OGAHC. These clusters correspond to subareas of the domain's state space and the connections between these subparts form bottlenecks of the domain. Learning skills which allow traversing these bottlenecks is a common approach to skill discovery in discrete domains (compare section 2.2).

3.2.1. Incremental transition graph estimation in continuous domains

A transition graph $G = (V, E, w)$ can be seen as a model of the domain's 1-step state transition probability (the domain's "dynamics"), where the nodes $v \in V$ represent "typical" states of the domain and edges $(v, v')_a \in E$ represent possible transitions in the domain under a specific action a . The edge weights w encode the corresponding probabilities $P_{vv'}^a$. In a model-free setting, G needs to be learned from experience. While this is straightforward in domains with discrete state space, it is more challenging in continuous domains. Force-based Iterative Graph Estimation (FIGE) is an heuristic approach to this problem with a solid theoretical motivation. FIGE learns transition graphs of size v_{num} from a set of state transitions $T = \{(s_i, a_i, s'_i)\}_{i=1}^n$ that have been experienced by the agent while acting in the domain. The transition graph is considered to be a generative model of state transitions and FIGE aims at finding graph node positions V which maximizes the likelihood of the observed transition (Metzen, in press).

FIGE is summarized in **Algorithm 1**: the set of graph nodes V with cardinality $|V| = v_{\text{num}}$ is initialized such that it covers the

Algorithm 1 | Force-based Iterative Graph Estimation (FIGE)

```

1: Input:  $T = \{(s_i, a_i, s'_i)\}_{i=1}^n$ , parameters  $v_{\text{num}}, K$ 
2:  $V = \text{INITIALIZE}(T, v_{\text{num}})$ 
3: For  $i = 0$  to  $K - 1$  do
4:   for all  $v \in V$  do
5:      $S^V(v) = \{s \mid (s, a, s') \in T : \text{NN}_V(s) = v\}$ 
6:      $F_S[v] = \text{MEAN}(S^V(v)) - v$ 
7:      $T^\rightarrow(v) = \{\text{NN}_V(s') - s + s \mid (s, a, s') \in T : \text{NN}_V(s) = v\}$ 
8:      $T^\leftarrow(v) = \{\text{NN}_V(s) - s + s' \mid (s, a, s') \in T : \text{NN}_V(s') = v\}$ 
9:      $F_G[v] = 0.5 \cdot [\text{MEAN}(T^\rightarrow(v)) + \text{MEAN}(T^\leftarrow(v))] - v$ 
10:    end for
11:     $V = V + \alpha_i \cdot 0.5(F_S[V] + F_G[V])$ 
12:  end for
13:  $N_{vv'}^a = |\{(s, s') \mid \exists (s, a, s') \in T : \text{NN}_V(s) = v \wedge \text{NN}_V(s') = v'\}|$ 
14:  $E = \{(v, v')_a \mid v, v' \in V \text{ } a \in A : N_{vv'}^a > 0\}$ 
15:  $w_{vv'}^a = N_{vv'}^a / \sum_{\tilde{v}} N_{v\tilde{v}}^a$ 

```

set of states contained in T uniformly by, e.g., maximizing the distance of the closest pair of graph nodes (line 2). Afterwards, for K iterations, the graph nodes are moved according to two kind of "forces" that act on them: the "sample representation" force (lines 5, 6) pulls each graph node v to the mean of all states S^V for which it is responsible, i.e., the states s for which it is the nearest neighbor $\text{NN}_V(s)$ in V . Thus, this force corresponds to an intrinsic k-means clustering of the observed states. The "graph consistency" force (lines 7–9) pulls each graph node v to a position where for all $(s, a, s') \in T$ with $\text{NN}_V(s) = v$ there is a vertex v' such that $v' - v$ is similar to $s' - s$, i.e., both vectors are close to parallel. Thus, this force encourages node positions which can represent the domain's dynamics well. The nodes are then moved according to the two forces (line 11), where the parameter $\alpha_i \in (0, 1]$

controls how greedily the node is moved to the position where the forces would become minimal. In order to ensure convergence of the graph nodes, α_i should go to 0 for i approaching K . An edge labeled with action a is added between two nodes v and v' if there exists at least one transition $(s, a, s') \in T$ with v being the nearest neighbor of s in V and v' being the nearest neighbor of s' in V (line 14). Furthermore, the edge weights are chosen as the empirical transition probabilities $\hat{P}_{vv'}^a$ from node v to v' under action a (line 15). For details and a derivation of FIGE, we refer to Metzen (in press).

The main drawbacks of FIGE are that the number of nodes of the transition graph need to be pre-specified and that FIGE is a batch algorithm and thus not well suited for incremental skill discovery. We present now Incremental FIGE (IFIGE) which does not suffer from these problems. IFIGE updates the graph's node positions after every experienced transition. Furthermore, IFIGE stores for every graph node v a set of exemplar states $S_v = \{s_i \mid i = 1, \dots, n_v\}$ and exemplar transitions $T_v = \{(s_i, a_i, s'_i) \mid i = 1, \dots, n_v\}$, with all s_i being "similar" to v and n_v being set typically to 25.

IFIGE starts with a single graph node $V = \{s_0\}$ and $S_{s_0} = T_{s_0} = \emptyset$, where s_0 is the start state. For any encountered transition (s, a, s') , the most similar graph node $v = \text{NN}_V(s)$, i.e., the nearest neighbor of s in V , is determined, s is added to the set of state exemplars S_v , and (s, a, s') to T_v . If the size of S_v or T_v exceeds n_v , old exemplars are deleted. Afterwards, the position of vertex v is updated using lines 5–9 of **Algorithm 1** for $T = T_v$. This changes the position of v ; thus, IFIGE checks afterwards for all state exemplars in S_v and transition exemplars in T_v whether any other node in V would be a better representative and moves the exemplars if required. Afterwards, IFIGE checks whether v is responsible for a too large area of the state space by computing the distance of the farthest pair in S_v . If this distance is above a threshold ζ , v is removed from V and two new nodes v_1 and v_2 are added to V . v_1 and v_2 are chosen as the cluster centers of a k -means clustering of S_v for $k = 2$. S_v and T_v are split into two subsets accordingly. Splitting nodes ensures that the number of graph nodes grows with the size of the state space explored by the agent.

When the current transition graph needs to be generated for skill discovery, IFIGE adds for all graph nodes v and any transition $(s, a, s') \in T_v$ an edge between v and $v' = \text{NN}_{V \setminus \{v\}}(s')$ for action a . Edge weights are determined by counting the frequencies of edges from v to v' relative to all edges starting from v .

3.2.2. Online graph-based agglomerative hierarchical clustering

Based on the transition graph, we identify task-independent and thus reusable skills using "Online Graph-based Agglomerative Hierarchical Clustering" (OGAHC). We give a brief summary of OGAHC and discuss how it can be extended to continuous domains; for more details we refer to the original publication (Metzen, 2012). OGAHC identifies skills by computing a *partition* P^* of the nodes V of a given transition graph G with respect to a prespecified linkage criterion l . Formally:

$$P^* = \arg \min_{P \in \mathcal{P}(V)} |P| \quad \text{s.t.} \quad \max_{p_i \in P, q_i \subset p_i} l(p_i \setminus q_i, q_i) \leq \psi,$$

with $\mathcal{P}(V)$ being the set of all possible partitions of V and ψ being a threshold which controls the granularity of the partition, i.e., the number of elements of the partition (called "cluster"). The aim is thus to compute a partition of the graph nodes with minimal cardinality such that the linkage between any pair of clusters of the partition is small, i.e., below ψ . Since this problem is \mathcal{NP} -hard, we use agglomerative hierarchical clustering as proposed by Mannor et al. (2004) for identifying an approximately optimal solution. As proposed by Simsek et al. (2005), we use the normalized cut \hat{N}_{cut} as linkage. The \hat{N}_{cut} of two disjoint subgraphs $A, B \subset G$ is an approximation of the probability that a random walk on G transitions in one time step from a state in subgraph A to a state in subgraph B or vice versa. Thus, we identify areas of the state space (corresponding to clusters of the graph) such that a randomly behaving agent would very unlikely leave one of these areas.

The connections of these clusters form *bottlenecks* of the graph and thus also of the domain. OGAHC creates one skill prototype for each pair of clusters $c_1, c_2 \in P^*$ which are connected in G ; this skill can be invoked any state s with $\text{NN}_V(s) \in c_1$ and terminates in any state with $\text{NN}_V(s) \notin c_1$. It terminates successfully if $\text{NN}_V(s) \in c_2$ and fails otherwise. Thus, the skill's objective is to guide the agent through one of the domain's bottlenecks from the area corresponding to cluster c_1 to the area of cluster c_2 .

Since the transition graph, which is the basis for OGAHC, is learned from experience and thus changes over time, performing the clustering only once is problematic: performing it early might result in a bad clustering of the domain since the transition graph might be inaccurate, while performing it late can overly increase the amount of experience the agent requires for skill discovery. Thus, it is desirable to perform the clustering several times during learning. For this, OGAHC assumes "dense local connectivity in the face of uncertainty," which prevents premature identification of bottlenecks and the corresponding skills, and adds constraints to the clustering process, which ensure that subsequent partitions remain consistent with prior ones. These constraints enforce that graph nodes that have been assigned to different clusters in one invocation of OGAHC remain in different clusters in later invocations.

The main hindrance of OGAHC in domains with continuous state space is that the constraints are based on the assumption that the graph nodes do not change over time. This is not the case when OGAHC is applied on top of IFIGE. This problem can be alleviated by adapting the current partition to the changes in the graph prior to any invocation of OGAHC. For this, let $P^*(V)$ be the partition of the graph nodes V of the last invocation of OGAHC and V' the current node positions. We extend $P^*(V)$ to a (pre-)partition P_{pre} of V' by assigning nodes $v'_a, v'_b \in V'$ to the same cluster if $\text{NN}_V(v'_a)$ and $\text{NN}_V(v'_b)$ are in the same cluster in $P^*(V)$. Now, OGAHC can be invoked with the usual constraints that nodes which are in different clusters in $P_{\text{pre}}(V')$ must be in different clusters in $P^*(V')$. For nodes $v' \in V'$ whose nearest neighbor $\text{NN}_V(v')$ is very different from v' , this constraint is relaxed, i.e., these nodes can be assigned to any cluster in $P^*(V')$. This corresponds to a situation where the agent has visited a particular area of the state space for the first time and the prior invocations of OGAHC put no restrictions on the bottlenecks in this novel part.

3.3. INTRINSIC MOTIVATION

In the context of this paper, intrinsic motivation refers to the process of mapping a transition from state s under option o to successor state s' onto an intrinsic reward r_i . We investigate two different intrinsic motivation mechanisms, one based on the *novelty* of a state under a skill and one based on the *prediction error* of a learned skill model.

For the novelty based motivation criterion, the agent stores for each option o the states it has encountered under this option so far in the set S_o^1 . When transitioning to state s' under option o , the intrinsic reward is computed via

$$r_i = - \sum_{j \in \text{NN}_{S_o}^{10}(s')} \exp\left(-\frac{\|s' - s_j\|_2^2}{b^2}\right),$$

where $\text{NN}_{S_o}^{10}(s')$ denotes the indices of the 10-nearest neighbors of s' in S_o and b is a domain-dependent scale parameter. Thus, the intrinsic reward is upper-bounded by 0 with values close to 0 if the 10 nearest neighbor are very different (large euclidean distance) from s' and very small values when s' is similar to several states in S_o . Thus, the novelty criterion discourages to execute options in regions of the state space where this option has been executed already several times. This mechanism is similar to the mechanism proposed by Hester and Stone (2012); however, in contrast to their work, it is also suited for domains with continuous state spaces.

For the prediction error criterion, the agent learns for each option a model \hat{P}_o that predicts the successor state of states s when following option o . The intrinsic reward is determined based on the error of the model's prediction via

$$r_i = -1 + \tanh(\sigma \|s' - \hat{P}_o(s)\|_2^2),$$

where σ is a domain-dependent scale parameter. The intrinsic reward r_i is large (close to 0) when the difference of predicted successor $\hat{P}_o(s)$ and actual successor s' is large. The intrinsic reward becomes small (close to -1) when the model correctly predicts the effect of executing option o in state s . Thus, the prediction error criterion encourages to execute options whose effects are unknown or unpredictable in the current area of the state space. Note that in contrast to the novelty criterion, for the prediction error criterion the intrinsic reward in a state depends on the option's policy.

The option model \hat{P}_o stores internally a set $T_o = \{(s_j, s'_j)\}$ of transitions encountered under option o . The model's prediction is based on 10-nearest neighbors regression:

$$P_o(s) = s + \frac{1}{10} \sum_{j \in \text{NN}_{T_o}^{10}(s)} (s'_j - s_j),$$

¹In order to keep the size of S_o limited, we remove states from S_o once $|S_o| > 2500$. The heuristic for selecting the state that is removed is to remove one of the states of the (approximate) closest state pair in S_o . This results in covering the effective state space of the problem approximately uniform.

where $\text{NN}_{T_o}^{10}(s)$ denotes the indices of the 10-nearest neighbors of s in the start states in T_o . If the size of T_o exceeds a threshold (in the experiments 2500) and a transition from s to s' is added, the oldest transition among $\text{NN}_{T_o}^{10}(s)$ is removed. This is required to keep the memory consumption limited and, more importantly, to track the non-stationarity in the target function that is induced by learning the option o concurrently and thus changing o 's policy.

4. RESULTS

In this section, we present an empirical evaluation of the proposed methods in two continuous and challenging RL benchmark domains. We evaluate both the behavior of the agent during the developmental period and its performance in external tasks. We have chosen these benchmark domains since they allow other researchers to compare their methods easily to our results.

4.1. 2D MULTI-VALLEY

4.1.1. Problem domain

The 2D Multi-Valley environment (see **Figure 3**) is an extension of the basic mountain car domain. The car the agent controls is not restrained to a one-dimensional surface, however, but to a two-dimensional surface. This two-dimensional surface consists of $2 \times 2 = 4$ valleys, whose borders are at $(\pi/6 \pm \pi/3, \pi/6 \pm \pi/3)$. The agent observes four continuous state variables: the positions in the two dimensions (x and y) and the two corresponding velocities (v_x and v_y). The agent can choose among the four discrete actions northwest, northeast, southwest, southeast which add $(\pm 0.001, \pm 0.001)$ to (v_x, v_y) . In each time step, due to gravity $0.004 \cos(3x)$ is added to v_x and $0.004 \cos(3y)$ to v_y . The maximal absolute velocity in each dimension is restrained to 0.07. The four valleys correspond naturally to clusters of the domain since transitioning from one valley to the other is unlikely under random behavior, i.e., represents a bottleneck. Thus, we would expect that one skill is created for each combination of adjacent valleys.

4.1.2. Developmental period

During its developmental period, the agent can explore the domain freely while engaging in skill discovery and following its intrinsic motivations. Initially, the agent has only a single option

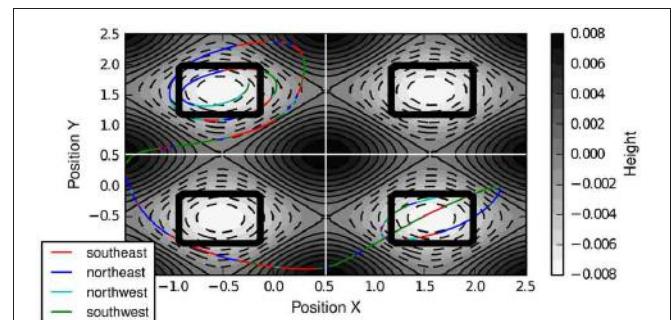


FIGURE 3 | 2D Multi-Valley domain. Gray-scale contours depict the height of the two-dimensional surface. The black boxes denote the target regions of the different tasks and the white lines the boundaries of the valleys. Shown is one example trajectory with color-coded actions.

o_e in its skill pool O , which can be invoked in any state of the environment, i.e., $I_{o_e} = S$, and terminates with probability $\beta_{o_e}(s) = 0.05$. This option can be considered to be the agent's exploration option, which can always be invoked if the agent prefers to explore the environment over learning a specific skill. We set the greediness of IFIGE to $\alpha_i = 0.25$ and the split node distance to $\zeta = 0.3$. For OGAHC, we set the maximal linkage to $\psi = -0.075$ and performed skill discovery every 5000 steps.

Each option's value function has been represented by an CMAC function approximator consisting of 10 independent tilings with $7^2 \cdot 5^2$ tiles, where the higher resolutions have been used for the x and y dimensions. The pseudo-reward for each option's policy has been set to $r_o = -1$ for each step and $r_o = -1000$ if an option terminates unsuccessfully, i.e., leaves its initiation set I_o without reaching its goal cluster c_2 . Value functions have been initialized to -100 . For learning the higher-level policy π_i , a lower resolution of $5^2 \cdot 3^2$ tiles has been used and the value functions have been initialized to 0. The discounting factor has been set to $\gamma = 0.99$ and all policies were ϵ -greedy with $\epsilon = 0.01$. The value functions were learned using Q-Learning and updated only for currently active options with a learning rate of 1. The scale-parameters of the intrinsic motivation mechanisms have been set to $b = 0.1$ (novelty) and $\sigma = 10^4$ (prediction error). All parameters have been chosen based on preliminary investigations.

Figure 4 shows the transition graphs generated by IFIGE after 20,000, 30,000, and 50,000 developmental steps. The two-dimensional embeddings of the graphs have been determined using Isomap (Tenenbaum et al., 2000). The four valleys of the domain clearly correspond to four densely connected subgraphs of the transition graph. The figure also shows that it would be difficult to determine a single point in time at which skill discovery should be performed: for instance, are the valleys $(0, 1)$ and $(1, 0)$ explored sufficiently after 30,000 steps to perform graph clustering? Since skill discovery with OGAHC is incremental, i.e., can be performed several times during learning, this choice need not be made.

Figure 5 shows the success ratio, i.e., how often a skill reaches its goal cluster, of the skills discovered during the developmental period. Initially, skills are unlikely to reach their goal area, with success ratios of approximately 0.25. Under both intrinsic motivation systems, the agent invests time in learning skill policies and the success ratio increases to 0.7 for the prediction error and 0.8 for the novelty criterion after approximately 10^5 steps of development. Note that success ratios of 1.0 are not possible since

for some states in $s \in I_o$, there is no way of reaching the option's goal area without leaving the initiation set, e.g., when the agent is moving with high velocity in the direction of the wrong neighbor valley. A possible explanation for the different performance under the two motivational systems is given below.

Figure 6 shows the ratio of selecting the option o_e ("Exploration") or any of the other, discovered options in O ("Skill Learning") under the policy π_i for different intrinsic motivations. Initially, no skills have been discovered and the agent thus has to explore. Once the first skills have been discovered, the agent focuses onto learning these skills. Over time, as the skill policies converge, a better predictive model for these skills can be learned. Similarly, the more time is spent on learning a skill, the less novel states are encountered under this skill. Accordingly, both intrinsic motivation mechanisms reduce the ratio of skill learning and focus on exploration again in order to discover new skills. Note that at this point in time, there are no further skills to be discovered in this domain but this is unknown to the agent.

In general, the prediction error-based motivation chooses the exploration option more often and reduces skill learning more abruptly than the novelty criterion. This can be explained by the fact that the exploration policy changes more strongly over time and it is thus harder to learn a model of this option. Once the policies of the other skills have settled, they are chosen only rarely. However, the results in **Figure 5** suggest that this happens too early as the final "fine-tuning" of the skill policies is not

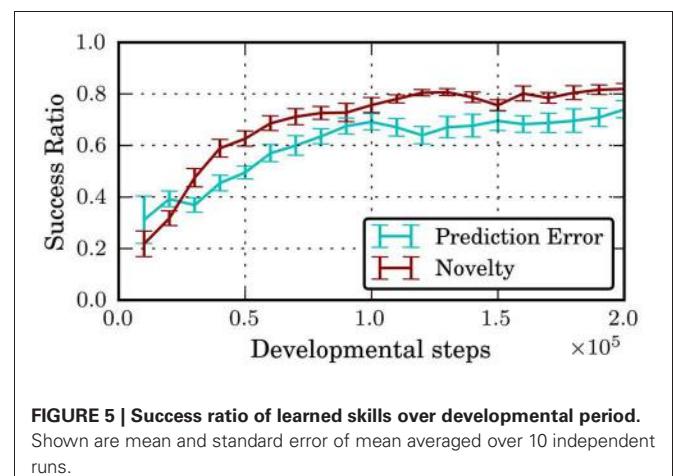


FIGURE 5 | Success ratio of learned skills over developmental period.
Shown are mean and standard error of mean averaged over 10 independent runs.

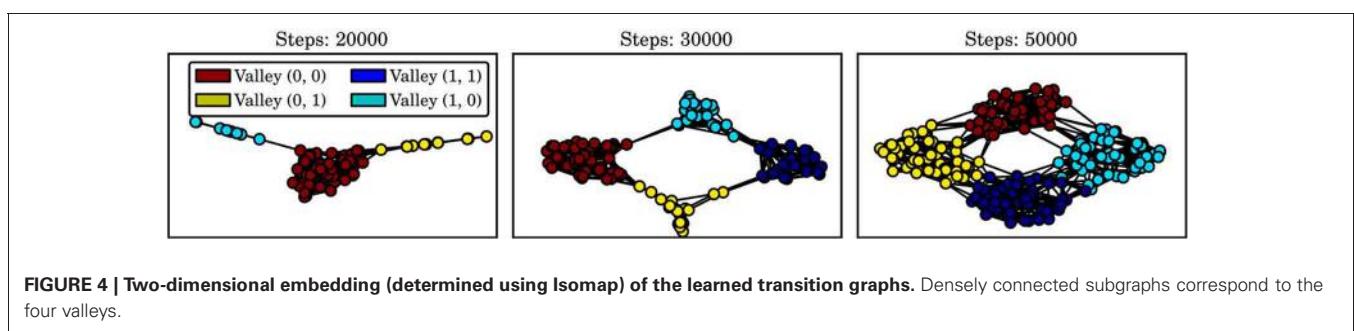


FIGURE 4 | Two-dimensional embedding (determined using Isomap) of the learned transition graphs. Densely connected subgraphs correspond to the four valleys.

finished and the success ratio is smaller than for the novelty criterion. Thus, the results indicate that using the prediction error for intrinsic motivation can be detrimental in situations where different option policies explore to different degrees since the prediction error criterion will favor the options with stronger exploration. Thus, it is recommended to base motivation on criteria like novelty or on the *change* of prediction error rather than on the error itself.

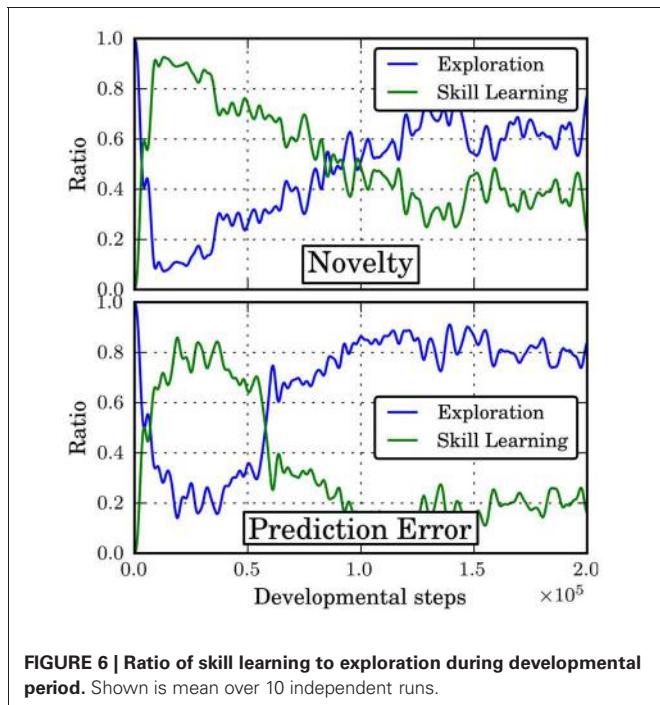


FIGURE 6 | Ratio of skill learning to exploration during developmental period. Shown is mean over 10 independent runs.

4.1.3. Task performance

In its “adulthood,” the agent is faced with a multi-task scenario: in each episode, the agent has to solve one out of 12 tasks. Each task is associated with a combination of two distinct valleys; e.g., in task (0, 1) the agent starts in the floor² of valley 0 and has to navigate to the floor of valley 1 and reduce its velocity such that $\|(v_x, v_y)\|_2 \leq 0.03$. In each time step, the agent receives an external reward of $r_e = -1$. Once a task is solved, the next episode starts with the car remaining at its current position and one of the tasks that starts in this valley is drawn at random. Episodes have been interrupted after 10^4 steps without solving the task and a new task was chosen at random. The current task is communicated as an additional state space dimension to the agent. The agent uses this task information and the reward r_e for learning the task policy π_e but ignores those information when improving π_o such that skills remain reusable in different tasks. The exploration option o_e used in the developmental period was removed from the skill set O such that the agent can only choose among self-discovered skills.

Figure 7 shows the results for different intrinsic motivation mechanisms and different lengths of the developmental period. As baseline, “No Skills” shows the performance of an agent that learns a monolithic policy for each task separately. For a very short developmental period of 10,000 steps, the hierarchical agent, which uses skills learned in the developmental period, learns initially faster than the monolithic agent, however, it converges to considerably worse policies. This is probably due to the fact that not all relevant skills have been discovered in the developmental period. See Jong et al. (2008) for a discussion of why an incomplete set of skills might have a detrimental effect

²The floor of valley 0 (see **Figure 3**) corresponds to the region $((-1/6 \pm 2/15)\pi, (-1/6 \pm 2/15)\pi)$.

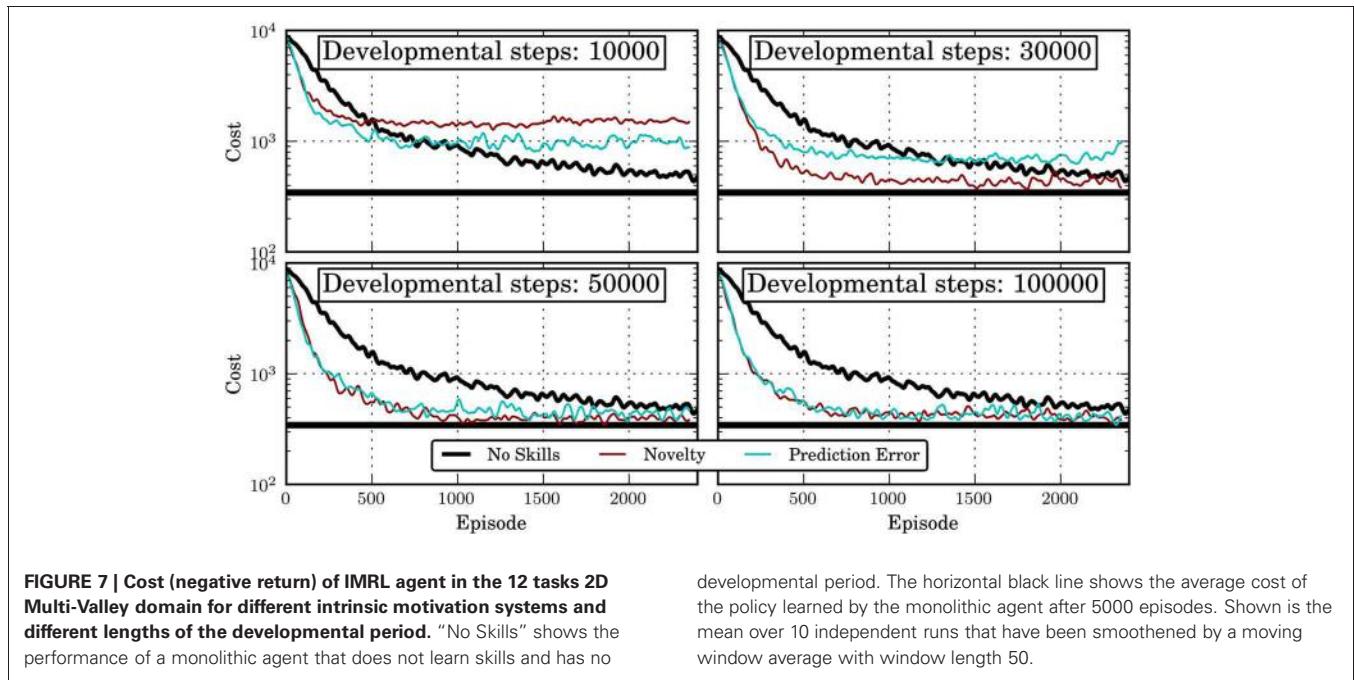


FIGURE 7 | Cost (negative return) of IMRL agent in the 12 tasks 2D Multi-Valley domain for different intrinsic motivation systems and different lengths of the developmental period. “No Skills” shows the performance of a monolithic agent that does not learn skills and has no

developmental period. The horizontal black line shows the average cost of the policy learned by the monolithic agent after 5000 episodes. Shown is the mean over 10 independent runs that have been smoothed by a moving window average with window length 50.

on an agent's performance. For 30,000 developmental steps, the skills acquired under the novelty motivation allow already to achieve close-to-optimal performance while the ones from the prediction-error motivation do not. This corresponds to the different qualities of the learned skills under the two motivation systems (compare **Figure 5**). For 50,000 or more developmental steps, the performance of the hierarchical agent approaches the optimal performance considerably faster than the monolithic agent, irrespective of the intrinsic motivation system used. This is interesting since after 50,000 steps, the learned skills are far from optimal (compare **Figure 5**). Apparently, also skills with sub-optimal policies can help the agent considerably. It should also be noted that even though a close-to-optimal performance is reached relatively fast, the performance remains slightly below the optimum which is reached by the monolithic agent after 5000 episodes. This is probably due to the (temporal) abstraction introduced by the skills which on the one hand helps the agent in learning faster but on the other hand also reduces the class of representable policies.

4.2. OCTOPUS

4.2.1. Problem domain

In the octopus arm domain³ (Yekutieli et al., 2005), the agent has to learn to control an Octopus arm. The base of the arm is restricted and cannot be actuated directly. The agent may control the arm in the following way: elongating or contracting the entire arm, bending the first half of the arm in either of the two directions, and bending the second half of the arm in either of the two directions. In each time step, the agent can set the elongation and the bending of the first and second half of the arm to an arbitrary value in $[-1, 1]$, resulting in 3 continuous action dimensions. The agent observes the positions x_i, y_i and velocities \dot{x}_i, \dot{y}_i of 24 selected parts of its arm (denoted by small black dots in **Figure 8**) and the angle and angular velocity of the arm's base. Thus, the state space is continuous and consists of 98 dimensions. Because of the high-dimensional and continuous state and action spaces and the complex dynamics of the domain, the octopus arm problem is a challenging task. It can also be seen as an easy simulation-based benchmark for actual robotic manipulation tasks.

³Source code available via <http://cs.mcgill.ca/dprecup/workshops/ICML06/octopus.html>.

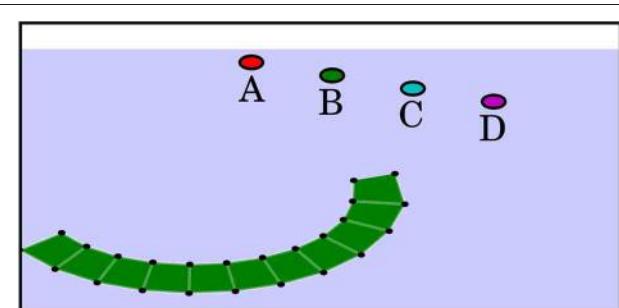


FIGURE 8 | Visualization of the octopus arm task. The circles represent target objects used in different tasks which yield an external reward when touched.

4.2.2. Developmental period

Similar to the developmental period in the 2D multi-valley domain, the agent can explore the domain freely while engaging in skill discovery and following its intrinsic motivations. However, the basis for skill discovery is not to identify bottlenecks (there are no bottlenecks in this domain) but to cluster the transition graph into regions which correspond to similar qualitative states. Thus, a different linkage criterion l_G has been used: for two sub-graphs A and B of the transition graph G , the linkage is set to $l_G(A, B) = 1/|A \cup B|^2 \sum_{v, v' \in A \cup B} d_{sp}(v, v')$, i.e., the average length of the shortest paths d_{sp} between two nodes in $A \cup B$. This linkage results in clusters with similar states in the sense that the agent can traverse from one state of the cluster to the other with a small number of steps. The maximum linkage ψ of a cluster in OGAHC has been set to 3.0 and skill discovery with OGAHC was performed every 10,000 steps. The greediness of IFIGE has been set to $\alpha_t = 0.25$ and the split node distance to $\zeta = 7.5$. Intrinsic motivation was based on the novelty mechanism with $b = 1$ and the length of the developmental period was set to 50,000 steps.

Because of the continuous action space, we have used direct policy search based on evolutionary computation for learning option policies π_o . The value for j -the action dimension is determined via $a_j = \tanh(\sum_{k=0}^{98} w_{jk} s_k)$, where s_k is the value of the k -th state dimension and $s_{98} = 1$ is a bias. The policy's weights w_{jk} have been optimized using 16 + 40 evolution strategy (ES) and each weight vector has been evaluated 10 times. The pseudo-reward for each option's policy has been set to $r_o = -1$ for each step and $r_o = -100$ if an option terminates unsuccessfully. The ES' objective is to maximize the pseudo-reward accumulated in 10 steps, after which the option is interrupted.

As in the multi-valley domain, the agent has initially only a single option o_e in its skill pool O , which can be invoked in any state of the environment, i.e., $I_{o_e} = S$, and terminates with probability $\beta_{o_e}(s) = 0.1$. π_{o_e} selects actions uniform randomly from the action space. The higher-level policy π_i , which determines the option that is executed, has been learned using Q-Learning with discounting factor $\gamma = 0.99$ and exploration rate $\epsilon = 0.01$. Because of the high dimensionality of the state space, the value function was not represented using a CMAC function approximator but using a linear combination of state values, i.e., $Q(s, o) = \sum_{k=0}^{98} w_{ok} s_k$. The learning rate has been set to 0.1.

4.2.3. Task performance

Different tasks can be imposed onto the agent; in this work, we require that the agent learns to reach for certain objects that are located at different positions (compare **Figure 8**). The agent obtains an external reward of -0.01 per time step and a reward of 100 for reaching the target object. The episode ends after 1000 time steps or once the target object is reached.

Figure 9 depicts an example trajectory of the octopus arm learned by the IMRL agent for reaching a target located at position C: the goal is reached after 22 steps and the agent invokes three different skills during this trajectory. The skill executed in the first 11 steps contracts the arm and brings it into an \cap -shape. The skill chosen for the next 6 steps unrolls the first part of the arm until an S-shape is reached. The skill executed in the last 5 steps unrolls the second half of the arm such that the target object is reached by

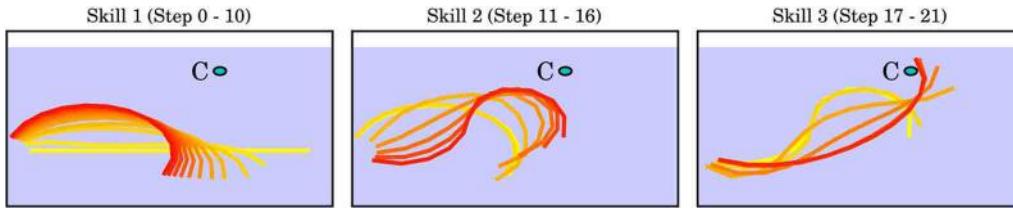


FIGURE 9 | Example trajectory of the octopus arm controlled by the IMRL agent. The trajectory corresponds to a sequence of three skills. Yellowish colored arms correspond to states at the beginning of skill execution while reddish colored arms correspond to states at the end of skill execution.

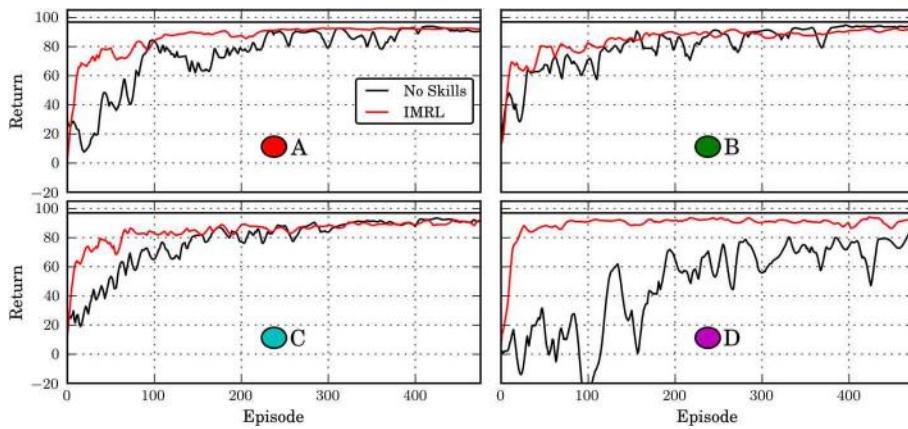


FIGURE 10 | Return of IMRL agent in the Octopus domain under the “novelty” motivation after 50,000 developmental steps. The circle patches indicate the respective targets used in the runs (compare Figure 8). “No Skills” shows the performance of an agent that does not learn skills and

has no developmental period. The horizontal black line shows the average cost of the policy learned by the monolithic agent after 2500 episodes. All curves show median performance over 5 independent runs and have been smoothed by a moving window average with window length 25.

an U-shape. Note that bending the arm directly into an U-shape would not be successful but result in a state like the one depicted in Figure 8.

Figure 10 shows the learning curves of the IMRL agent and a monolithic agent, which learns a flat global policy with the same parametrization as the skill policies, for different target positions in the Octopus domain. Given sufficient time, the monolithic agent can learn policies of similar quality as the IMRL agent. Thus, close-to-optimal behavior can be represented by a flat global policy. However, in general, the IMRL agent learns close-to-optimal policies faster and the learning curves exhibit less variance across all tasks. Thus, the temporal abstraction of the skills that were learned in the developmental period seem to make learning close-to-optimal behavior easier by providing a useful explorative bias. On the other hand, as in the multi-valley domain these abstractions may impair performance slightly in the long run.

5. CONCLUSION AND FUTURE WORK

We have presented a novel skill discovery approach suited for continuous domains that can be used by an IMRL agent in its developmental period. Our empirical results in two continuous RL domains suggest that the IMRL agent benefits from the discovered skills once it is faced with external tasks: close-to-optimal behaviors can be learned in less trials because of the explorative bias provided by the temporal abstractions of

the skill hierarchy. However, this explorative bias is only helpful if the developmental period was sufficiently long: if the learning and discovery of skills is interrupted prematurely, an IMRL agent might perform worse than an agent which learns a monolithic policy from scratch. Furthermore, we have compared two intrinsic motivation mechanisms and presented evidence that intrinsic motivation allows to reasonably determine how much time should be spent on learning specific skills.

This work can be extended in numerous ways: for instance, instead of performing skill discovery only in the developmental period, the agent could also discover novel skills and learn based on intrinsic motivation while he is faced with an external task. This, however, requires trading off intrinsic and external rewards and facing the exploration-exploitation dilemma. We leave this to future work; however, we would like to emphasize that the proposed skill discovery approach is in no way restricted to the developmental setting. A further direction of future work would be to combine the proposed skill discovery approach with more sophisticated intrinsic motivation mechanisms such as competence progress intrinsic motivation (Stout and Barto, 2010) or other means for empirically estimating the learning progress (see, e.g., Lopes et al., 2012). Furthermore, it would be desirable to learn more complex hierarchies of skills, where skills can invoke other skills. The dendrogram generated by the hierarchical clustering in OGABC could be an interesting starting

point for this. For being useful in a realistic robotic setup, the proposed methods would need to be integrated into a control architecture with, e.g., reactive behaviors and predictive control, such as the one shown in **Figure 1**. This should allow to deal better with non-markovian, noisy, and partial observable problems.

REFERENCES

- Baldassarre, G. (2011). "What are intrinsic motivations? A biological perspective," in *IEEE International Conference on Development and Learning*. Vol. 2 (Frankfurt am Main), 1–8. doi: 10.1109/DEVLRN.2011.6037367
- Barto, A. G., and Mahadevan, S. (2003). Recent advances in hierarchical reinforcement learning. *Dis. Event Dyn. Syst.* 13, 341–379. doi: 10.1023/A:1022140919877
- Barto, A. G., Singh, S., and Chentanez, N. (2004). "Intrinsically motivated learning of hierarchical collections of skills," in *Proceedings of the 3rd International Conference of Developmental Learning* (LaJolla, CA), 112–119. doi: 10.1.1.117.6436
- Harlow, H. F. (1950). Learning and satiation of response in intrinsically motivated complex puzzle performance by monkeys. *J. Compar. Physiol. Psychol.* 43, 289–294. doi: 10.1037/h0058114
- Hester, T., and Stone, P. (2012). "Intrinsically motivated model learning for a developing curious agent," in *Proceedings of the 11th International Conference on Development and Learning* (San Diego, CA). doi: 10.1109/DevLrn.2012.6400802
- Jong, N. K., Hester, T., and Stone, P. (2008). "The utility of temporal abstraction in reinforcement learning," in *Proceedings of the 7th Conference on Autonomous Agents and Multiagent Systems* (Estoril), 299–306.
- Kirchner, F. (1998). Q-learning of complex behaviours on a six-legged walking machine. *J. Robot. Auton. Syst.* 25, 256–263. doi: 10.1016/S0921-8890(98)00054-2
- Kirchner, F., and Richter, C. (2000). "Q-surfing: exploring a world model by significance values in reinforcement learning tasks," in *Proceedings of the European Conference on Artificial Intelligence* (Berlin), 311–315.
- Köhler, T., Rauch, C., Schröer, M., Berghöfer, E., and Kirchner, F. (2012). "Concept of a biologically inspired robust behaviour control system," in *Proceedings of 5th International Conference on Intelligent Robotics and Applications* (Montreal, QC), 486–495. doi: 10.1007/978-3-642-33515-0_48
- Konidaris, G., and Barto, A. G. (2009). "Skill discovery in continuous reinforcement learning domains using skill chaining," in *Advances in Neural Information Processing Systems (NIPS)*. Vol. 22 (Vancouver, BC), 1015–1023.
- Lopes, M., Lang, T., Toussaint, M., and Oudeyer, P.-Y. (2012). "Exploration in model-based reinforcement learning by empirically estimating learning progress," in *Advances in Neural Information Processing Systems (NIPS)* (Lake Tahoe, Nevada), 206–214.
- Mannor, S., Menache, I., Hoze, A., and Klein, U. (2004). "Dynamic abstraction in reinforcement learning via clustering," in *Proceedings of the 21st International Conference on Machine Learning* (Banff, AB), 560–567. doi: 10.1145/1015330.1015355
- McGovern, A., and Barto, A. G. (2001). "Automatic discovery of subgoals in reinforcement learning using diverse density," in *Proceedings of the 18th International Conference on Machine Learning* (Williamstown, MA), 361–368.
- Menache, I., Mannor, S., and Shimkin, N. (2002). "Q-Cut – dynamic discovery of sub-goals in reinforcement learning," in *Proceedings of the 13th European Conference on Machine Learning* (Helsinki, Finland), 295–306. doi: 10.1007/3-540-36755-1_25
- Metzen, J. H. (2012). Online skill discovery using graph-based clustering. *J. Mach. Learn. Res. W&CP* 24, 77–88.
- Metzen, J. H. (in press). "Learning graph-based representations for continuous reinforcement learning domains," in *Proceedings of the European Conference on Machine Learning (ECML 2013)*, (Prague: Springer).
- Oudeyer, P.-Y., and Kaplan, F. (2007). What is intrinsic motivation? A typology of computational approaches. *Front. Neurorobot.* 1:6. doi: 10.3389/neuro.12.006.2007
- Oudeyer, P.-Y., Kaplan, F., and Hafner, V. (2007). Intrinsic motivation systems for autonomous mental development. *IEEE Trans. Evol. Comput.* 11, 265–286. doi: 10.1109/TEVC.2006.890271
- Randløv, J., and Alstrøm, P. (1998). "Learning to drive a bicycle using reinforcement learning and shaping," in *Proceedings of the 15th International Conference on Machine Learning* (Madison, WI), 463–471.
- Schembri, M., Mirolli, M., and Baldassarre, G. (2007). "Evolution and learning in an intrinsically motivated reinforcement learning robot," in *Proceedings of the 9th European Conference on Advances in Artificial Life* (Lisbon, Portugal), 294–303. doi: 10.1007/978-3-540-74913-4_30
- Schmidhuber, J. (1991). "Curious model-building control systems," in *Proceedings of the International Joint Conference on Neural Networks* (Singapore: IEEE), 1458–1463.
- Şimşek, Ö., and Barto, A. G. (2004). "Using relative novelty to identify useful temporal abstractions in reinforcement learning," in *Proceedings of the 21st International Conference on Machine Learning* (Banff, AB), 751–758. doi: 10.1145/1015330.1015353
- Şimşek, Ö., and Barto, A. G. (2009). "Skill characterization based on betweenness," in *Advances in Neural Information Processing Systems (NIPS)*. Vol. 22 (Vancouver, BC), 1497–1504.
- Şimşek, Ö., Wolfe, A. P., and Barto, A. G. (2005). "Identifying useful subgoals in reinforcement learning by local graph partitioning," in *Proceedings of the 22nd International Conference on Machine Learning* (Bonn, Germany), 816–823. doi: 10.1145/1102351.1102454
- Skinner, B. (1938). *The Behavior of Organisms: An Experimental Analysis*. The Century Psychology Series. New York, NY: Appleton-Century-Crofts.
- Stout, A., and Barto, A. G. (2010). "Competence progress intrinsic motivation," in *Proceedings of the 9th IEEE International Conference on Development and Learning* (Ann Arbor, MI), 257–262. doi: 10.1109/DEVLRN.2010.5578835
- Sutton, R. S., and Barto, A. G. (1998). *Reinforcement Learning: An Introduction*. Cambridge, MA: The MIT Press.
- Sutton, R. S., Koop, A., and Silver, D. (2007). "On the role of tracking in stationary environments," in *Proceedings of the 24th International Conference on Machine Learning* (Corvallis, OR: ACM), 871–878. doi: 10.1145/1273496.1273606
- Sutton, R. S., Precup, D., and Singh, S. (1999). Between MDPs and semi-MDPs: a framework for temporal abstraction in reinforcement learning. *Artif. Intell.* 112, 181–211. doi: 10.1016/S0004-3702(99)00052-1
- Tenenbaum, J. B., Silva, V. D., and Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science* 290, 2319–2323. doi: 10.1126/science.290.5500.2319
- Thrun, S. (1996). "Is learning the n-th thing any easier than learning the first?" in *Advances in Neural Information Processing Systems (NIPS)* (Cambridge, MA: MIT Press), 640–646.
- Whiteson, S. (2012). "Evolutionary computation for reinforcement learning," in *Reinforcement Learning: State of the Art* (Berlin: Springer), 325–358.
- Yekutieli, Y., Sagiv-Zohar, R., Aharonov, R., Engel, Y., Hochner, B., and Flash, T. (2005). A dynamic model of the octopus arm. I. Biomechanics of the octopus reaching movement. *J. Neurophysiol.* 5, 291–323.

ACKNOWLEDGMENTS

This work was supported through a grant of the German Federal Ministry of Economics and Technology (BMWi, FKZ 50 RA 1217). The authors would like to thank Yohannes Kassahun for helpful comments on this work.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 21 May 2013; accepted: 10 July 2013; published online: 26 July 2013.

Citation: Metzen JH and Kirchner F (2013) Incremental learning of skill collections based on intrinsic motivation. *Front. Neurorobot.* 7:11. doi: 10.3389/fnbot.2013.00011

Copyright © 2013 Metzen and Kirchner. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in other forums, provided the original authors and source are credited and subject to any copyright notices concerning any third-party graphics etc.



Neural model for learning-to-learn of novel task sets in the motor domain

Alexandre Pitti*, Raphaël Braud, Sylvain Mahé, Mathias Quoy and Philippe Gaussier

ETIS Laboratory, UMR CNRS 8051, the University of Cergy-Pontoise, ENSEA, Cergy-Pontoise, France

Edited by:

Gianluca Baldassarre, Italian National Research Council, Italy

Reviewed by:

Vincenzo G. Fiore, UCL Institute of Neurology, UK

Masaki Ogino, Kansai University, Japan

***Correspondence:**

Alexandre Pitti, ETIS - UMR CNRS 8051, Université de Cergy-Pontoise, St-Martin 1 2, avenue Adolphe-Chauvin, Cergy-Pontoise F 95302, France

e-mail: alexandre.pitti@u-cergy.fr

During development, infants learn to differentiate their motor behaviors relative to various contexts by exploring and identifying the correct structures of causes and effects that they can perform; these structures of actions are called *task sets* or *internal models*. The ability to detect the structure of new actions, to learn them and to select on the fly the proper one given the current task set is one great leap in infants cognition. This behavior is an important component of the child's ability of learning-to-learn, a mechanism akin to the one of intrinsic motivation that is argued to drive cognitive development. Accordingly, we propose to model a dual system based on (1) the learning of new task sets and on (2) their evaluation relative to their uncertainty and prediction error. The architecture is designed as a two-level-based neural system for context-dependent behavior (the first system) and task exploration and exploitation (the second system). In our model, the task sets are learned separately by reinforcement learning in the first network after their evaluation and selection in the second one. We perform two different experimental setups to show the sensorimotor mapping and switching between tasks, a first one in a neural simulation for modeling cognitive tasks and a second one with an arm-robot for motor task learning and switching. We show that the interplay of several intrinsic mechanisms drive the rapid formation of the neural populations with respect to novel task sets.

Keywords: task sets, fronto-parietal system, decision making, incremental learning, cortical plasticity, error-reward processing, gain-field mechanism, tool-use

1. INTRODUCTION

The design of a multi-tasks robot that can cope with novelty and evolve in an open-ended manner is still an open challenge for robotics. It is however an important goal (1) for conceiving personal assistive robots that are adaptive (e.g., to infants, the elderly and to the handicapped people) and (2) for studying from an inter-disciplinary viewpoint the intrinsic mechanisms underlying decision making, goal-setting and the ability to respond on the fly and adaptively to novel problems.

For instance, robots cannot yet reach the level of infants for exploring alternative ways to surmount an obstacle, searching for a hidden toy in a new environment, finding themselves the proper way to use a tool, or solving a jigsaw puzzle. All these tasks require to be solved within boundaries of their given problem space, without exploring it entirely. Thus, robots lack this ability to detect and explore new behaviors and action sequences oriented toward a goal; i.e., what is called a *task set* (Harlow, 1949; Collins and Koehlein, 2012).

The ability to manipulate dynamically task sets is however a fundamental aspect of cognitive development (Johnson, 2012). Early in infancy, infants are capable to perform flexible decision-making and dynamic executive control even at a simple level in order to deal with the unexpected (Tenenbaum et al., 2011). Later on, when they are more mature, they learn to explore the tasks space, to select goals and to focus progressively on tasks of increasing complexity. One example in motor development is the learning of different postural configurations. Karen Adolph

explains for instance how infants progressively differentiate their motor behaviors into task sets (i.e., the motor repertoire) and explore thoroughly the boundaries of each postural behavior till becoming expert on what they discover (Adolph and Joh, 2005, 2009). Adolph further argues that the building of a motor repertoire is not preprogrammed with a specific developmental timeline but that each postural behavior can be learned independently as separated tasks without pre-ordered dependencies to the other ones (crawling, sitting, or standing).

This viewpoint is also shared by neurobiologists who conceive the motor system to structure the actions repertoire into "internal models" for each goal to achieve (Wolpert and Flanagan, 2010; Wolpert et al., 2011). Each novel contextual cue (e.g., handling a novel object) promotes the acquisition and the use of a distinct internal model that does not modify the existing neural representations used to control the limb on its own (White and Diedrichsen, 2013). Moreover, each task set is evaluated depending on the current dynamics and on the current goal we want to perform (Orban and Wolpert, 2011). For instance, we switch dynamically from different motor strategies to the most appropriate one depending on the context; e.g., tilting the racket to the correct angle in order to give the desired effect on the ball, or for executing the proper handling of objects with respect to their estimated masses (Cothros et al., 2006).

From a developmental viewpoint, the capability for flexible decision-making gradually improves in 18 months-old infants (Tenenbaum et al., 2011). Decision-making endows

infants to evaluate the different alternatives they have for achieving one goal with respect to the ongoing sequence and to select the correct one(s) among different alternatives. It owes them also the possibility to inhibit some previously learned strategies in order to explore new ones never seen before (Yokoyama et al., 2005).

IN AI, this craving to explore, to test and to embed new behaviors is known as intrinsic motivation (Kaplan and Oudeyer, 2007). In Kaplan and Oudeyer's words: "The idea is that a robot (...) would be able to autonomously explore its environment not to fulfill predefined tasks but driven by some form of intrinsic motivation that pushes it to search for situations where learning happens efficiently". In this paper, we focus more on the idea that the rewards are self-generated by the machine itself (Singh et al., 2010) and that the function of intrinsic motivation is mainly to regulate the exploration/exploitation problem, driving exploratory behavior and looking for different successful behaviors in pursuing a goal. In that context, we propose that the ability to choose whether or not to follow the same plan or to create a novel one out of nothing—in regard to the current situation—is an intrinsic motivation. We studied for instance the role of the neuromodulator acetylcholine in the hippocampus for novelty detection and memory formation (Pitti and Kuniyoshi, 2011).

Meanwhile, the capability to make decision and to select between many options is one important aspect of intrinsic motivation because otherwise the system would be only passive and would not be able to select or encourage one particular behavior. Taking decisions in deadlock situations requires therefore some problem-solving capabilities like means-end reasoning (Koechlin et al., 2003) and error-based learning capabilities (Adolph and Joh, 2009). For instance, means-end reasoning and error-based learning are involved in some major psychological tests such as the Piagetian "A-not-B error test" (Diamond, 1985; Smith et al., 1999; Schöner and Dineva, 2007), Harlow's learning set test (Harlow, 1949) and tool-use (Lockman, 2000; Fagard et al., 2012; Vaesen, 2012; Guerin et al., 2013). The A-not-B error test describes a decision-making problem where a 9-month old infant still pertains to select an automatic wrong response (e.g., the location A) and cannot switch dynamically from this erroneous situation to the correct one (e.g., the location B). Above this age, however, infants do not make the error and switch rapidly to the right location. A similar observation is found in Harlow's experiments on higher learning (Harlow, 1949) where Rhesus monkeys and humans have to catch the pattern of the experiment in a series of learning experiences. Persons and monkeys demonstrate that they learn to respond faster when facing a novel and similar situation by switching to the correct strategy, by catching the pattern to stop making the error: they show therefore that they do not master isolated tasks but, instead, they grasp the relation between the events. In one situation, if the animal guessed wrong on the first trial, then it should switch directly to the other solution. In another situation, if it guessed right on the first trial, then it should continue. This performance seems to require that the monkey, the baby or the person use an abstract rule and solve the problem with an apparent inductive reasoning (Tenenbaum et al., 2011). In line with these observations on the development of flexible behaviors, researchers focused on tool-use: when infants start to use an object as a means to an end, they serialize

their actions toward a specific goal, as for example reaching a toy with a stick (Fagard et al., 2012; Rat-Fischer et al., 2012; Guerin et al., 2013). Tool-use requires also finding patterns like the shape of grasping, order and sequentiality of patterns (Cothros et al., 2006).

Considering the mechanisms it may involve, Karen Adolph emphasizes the ability of *learning-to-learn* (Adolph and Joh, 2005), a process akin to Harlow (1949). Harlow coined the expression to distinguish the means for finding solutions to novel problems from simple associative learning and stimulus generalization (Adolph, 2008). Adolph reinterprets this proposal and suggests that two different kinds of thinking and learning are at work in the infant brain, governing the aspects of exploration and of generalization (Adolph and Joh, 2009). On the one hand, one learning system is devoted to the learning of task sets from simple stimulus-response associations. For instance, when an infant recognizes the context, he selects his most familiar strategy and reinforces it within his delimiting parameter ranges. On the other hand, a second learning is devoted to detect a new situation as is and to find a solution dynamically in a series of steps. Here, the acceptance of uncertainty gradually leads for making choices and decisions in situation never seen before. However, which brain regions and which neural mechanisms this framework underlies?

Among the different brain regions, we emphasize that the post-parietal cortex (PPC) and the pre-frontal cortex (PFC) are found important (1) for learning context-dependent behavior and (2) for evaluating and selecting these behaviors relative to their uncertainty and error prediction. Regarding the PPC, different sensorimotor maps co-exist to represent structured information like spatial information or the reaching of a target, built on coordinate transform mechanisms (Stricanne et al., 1996; Andersen, 1997; Pouget and Snyder, 2000). Furthermore, recent studies acknowledge the existence of context-specific neurons in the parieto-motor system for different grasp movements (Brozovic et al., 2007; Andersen and Cui, 2009; Baumann et al., 2009; Fluet et al., 2010). Regarding the PFC, Johnson identifies the early development of the pre-frontal cortex as an important component for enabling executive functions (Johnson, 2012) while other studies have demonstrated difficulty in learning set formation following extensive damage of the prefrontal cortex (Warren and Harlow, 1952; Yokoyama et al., 2005). The PFC manipulates information on the basis of the current plan (Fuster, 2001), and it is active when new rules need to be learned and other ones rejected. Besides, its behavior is strongly modulated by the anterior cingulate cortex (ACC) which plays an active role for evaluating task sets and for detecting errors during the current episode (Botvinick et al., 2001; Holroyd and Coles, 2002; Khamassi et al., 2011). If we look now at the functional organization of these brain structures, many authors emphasize the interplay between an associative memory of action selection in the temporal and parietal cortices (i.e., an integrative model) and a working memory for actions prediction and decision making in the frontal area (i.e., a serial model) (Fuster, 2001; Andersen and Cui, 2009; Holtmaat and Svoboda, 2009). All-in-all, these considerations permit us to draw a scenario based on a two complementary learning systems.

More precisely, we propose to model a dual system based on (1) the learning of task sets and on (2) the evaluation of

these task sets relative to their uncertainty, and error prediction. Accordingly, we design a two-level based neural system for context-dependent behavior (PPC) and task exploration and prediction (ACC and PFC); see **Figure 1**. In our model, the task sets are learned separately by reinforcement learning in the post parietal cortex after their evaluation and selection in the prefrontal cortex and anterior cingulate cortex. On the one hand, the learner or agent stores and exploits its familiar knowledge through a reinforcement learning algorithm into contextual patterns called and collected from all its different modalities. On the other hand, the learner evaluates and compares the way it learns, and selects the useful strategies while it discards others or tests new ones on the fly if no relevant strategy is found. We perform two different experimental setups to show the sensorimotor mapping and switching between tasks, one in a neural simulation for modeling cognitive tasks and another with an arm-robot for motor task learning and switching. We use neural networks to learn simple sensorimotor mapping for different tasks and compute their variance and error for estimating the sensorimotor prediction. Above a certain threshold, the error signal is used to select and to evaluate the current strategy. If no strategy is found pertinent for the current situation, this corresponds to a novel motor schema that is learned independently by a different map. In a cognitive experiment similar to Harlow (1949) and Diamond (1990), we employ this neural structure to learn multiple spatio-temporal sequences and switch between different strategies if an error has occurred or if a reward has been received (error-learning). In a psycho-physic experiment similar to Wolpert and Flanagan (2010), we show how a robotic arm learns the visuomotor strategies for stabilizing the end-point of its own arm when it moves it alone and when it is holding a long stick. Here, the uncertainty on the spatial location of the end-point triggers the decision-making from the two strategies by selecting the best one given the proprioceptive and visual feedback and the error signal delivered.

2. MATERIALS AND METHODS

In this section, we present the neural architecture and the mechanisms that govern the dynamics of the neurons, of reinforcement learning and of decision-making. We describe first the bio-inspired mechanism of rank-order coding from which we derive the activity of the parietal and of the pre-frontal neurons. In

second, we describe the reinforcement learning algorithm, the error prediction reward and the decision-making rules.

2.1. PPC—GAIN-FIELD MODULATION AND SENSORIMOTOR MAPPING

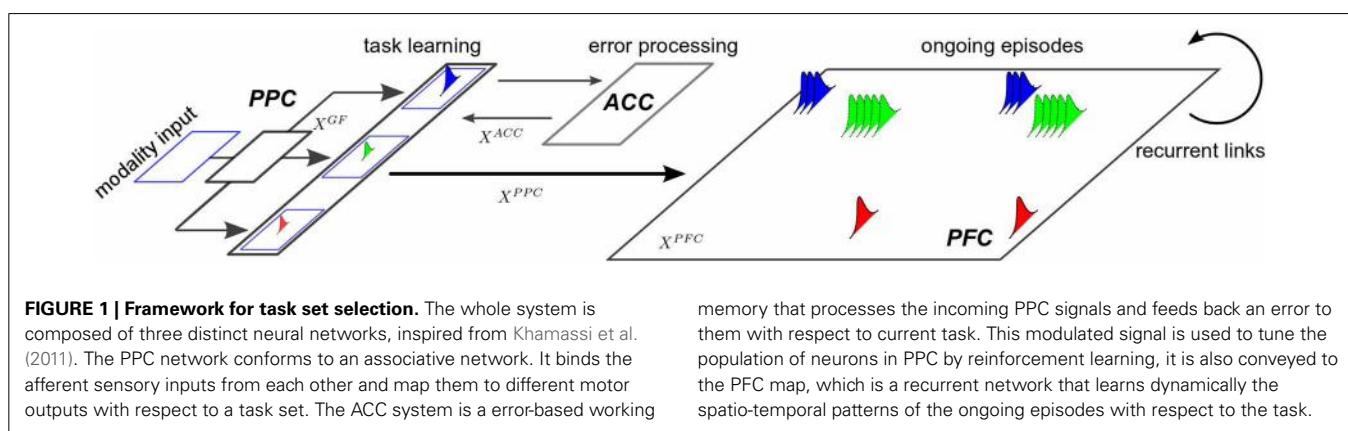
We employ the rank-order coding neurons to model the sensorimotor mapping between input and output signals with an architecture that we have used in a previous research (Pitti et al., 2012). This architecture implements multiplicative neurons, called gain-field neurons, that multiply unit by unit the value of two or more incoming neural populations, see **Figure 2**. Its organization is interesting because it transforms the incoming signals into a basis functions' representation that could be used to simultaneously represent stimuli in various reference frames (Salinas and Thier, 2000). The multiplication between afferent sensory signals in this case from two population codes, X_{m_1} and X_{m_2} , $\{m_1, m_2 \in M_1, M_2\}$, produces the signal activity X_n to the n gain-field neurons, $n \in N$:

$$X^{GF} = X^{M_1} \times X^{M_2} \quad (1)$$

The key idea here is that the gain-field neurons encode two information at once and that the amplitude of the gain-field neurons relates the values of one modality *conditionally* to the other; see **Figure 2A**. The task is therefore encoded into a space of lower dimension (Braun et al., 2009, 2010). We exploit this feature to model the parietal circuits for different contextual cues and internal models, which means that, after the encoding, the output layers learn the receptive fields of the gain-field map and translates this information into various gain levels. In **Figure 2B**, we give a concrete example of one implementation, here delineated to two modalities, with N gain-fields projecting to three different tasks set of different size. We explain thereafter (1) how the gain fields neurons learn the associations between various modalities and (2) how the neurons of the output map learn from the gain fields neurons for each desired task.

2.2. RANK-ORDER CODING ALGORITHM

We implement a hebbian-like learning algorithm proposed by Van Rullen et al. (1998) called the Rank-Order Coding (ROC) algorithm. The ROC algorithm has been proposed as a discrete and faster model of the derivative integrate-and-fire neuron (Van Rullen and Thorpe, 2002). ROC neurons are sensitive



to the sequential order of the incoming signals; that is, its *rank code*, see **Figure 3A**. The distance similarity to this code is transformed into an amplitude value. A scalar product between the input's rank code with the synaptic weights furnishes then a distance measure and the activity level of the neuron. More precisely, the ordinal rank code can be obtained by sorting the signals' vector relative to their amplitude levels or to their temporal order in a sequence. We use this property respectively for modeling the signal's amplitude for the parietal neurons and the spatio-temporal patterns for the prefrontal neurons. If the rank code of the input signal matches perfectly the one of the synaptic weights, then the neuron fully integrates this activity over time and fires, see **Figure 3A**. At contrary, if the rank order of the signal vector does not match properly the ordinal sequence of the synaptic weights, then integration is weak and the neuron discharges proportionally to it, see **Figure 3B**.

The neurons' output X is computed by multiplying the rank order of the sensory signal vector I , $\text{rank}(I)$, by the synaptic weights w ; $w \in [0, 1]$. For a vector signal of dimension M and for a population of N neurons (M afferent synapses), we have for the

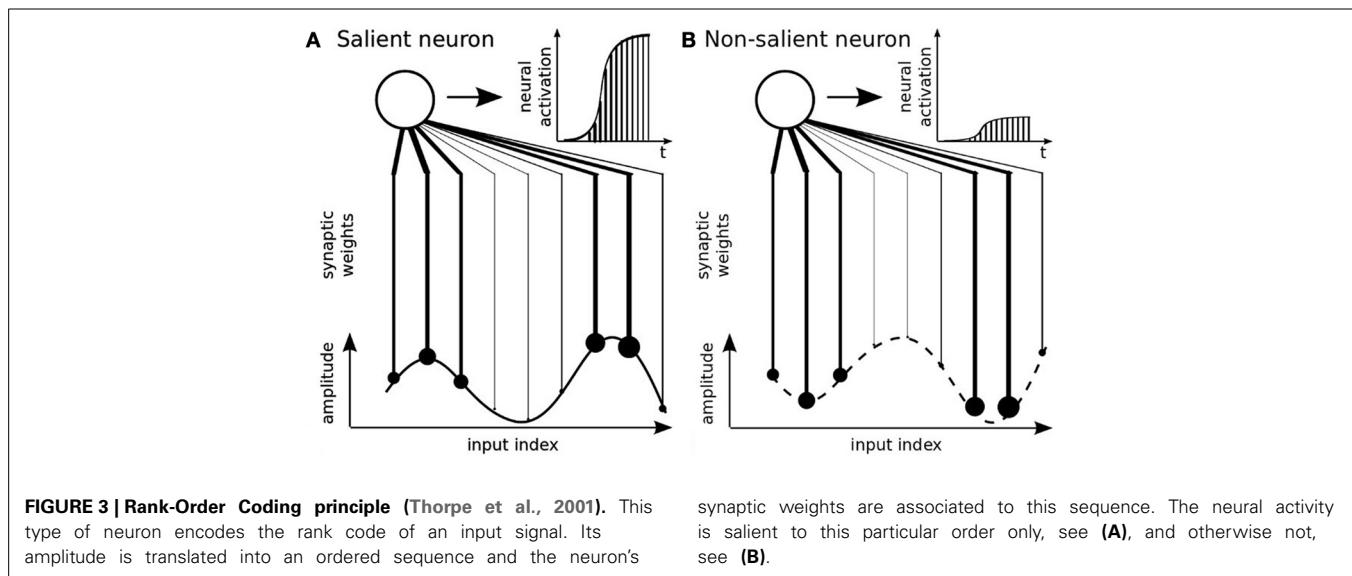
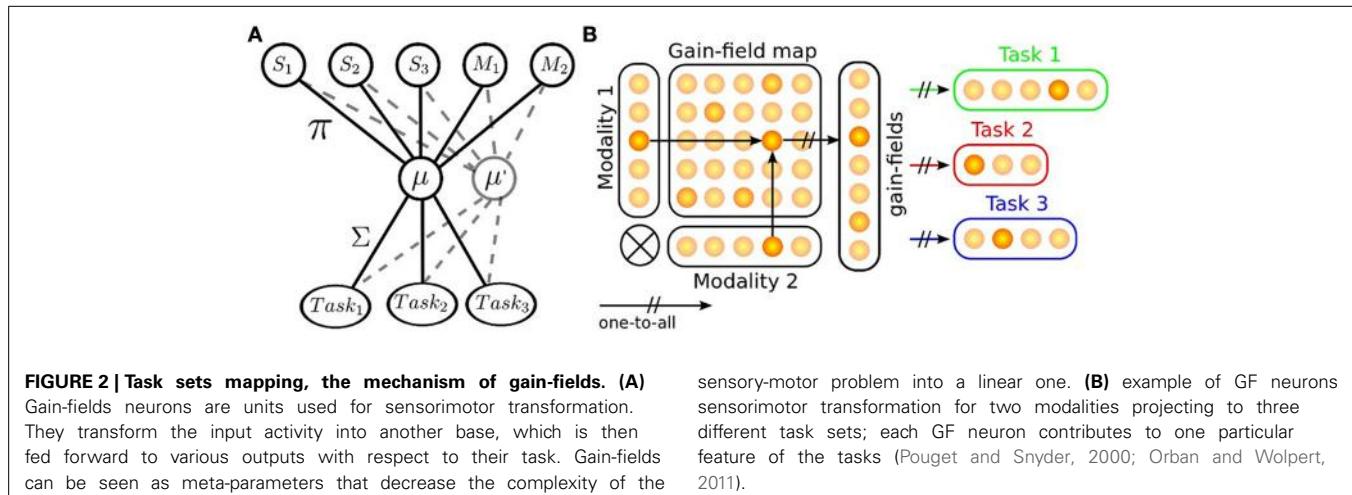
GF neurons and for the output PPC neurons:

$$\begin{cases} X_n^{GF} = \sum_{m \in M} \frac{1}{\text{rank}(I_m)} w_{n, m}^{GF-\text{Modality}} \\ X_n^{PPC} = \sum_{m \in M} \frac{1}{\text{rank}(I_m)} w_{n, m}^{PPC-GF} \end{cases} \quad (2)$$

The updating rule of the neurons' weights is similar to the winner-takes-all learning algorithm of Kohonen's self-organizing maps (Kohonen, 1982). For the best neuron $s \in N$ and for all afferent signals $m \in M$, we have for the neurons of the output layer:

$$\begin{cases} w_{s, m}^{PPC-GF} = w_{s, m}^{PPC-GF} + \Delta w_{s, m}^{PPC-GF} \\ \Delta w_{s, m}^{PPC-GF} = \frac{1}{\text{rank}(I_m)} - w_{s, m}^{PPC-GF}, \end{cases} \quad (3)$$

the equations are the same for GF neurons (not reproduced here). We make the note that the synaptic weights follow a power-scale density distribution that makes the rank-order coding neurons similar to basis functions. This attribute permits to use them as receptive fields so that the more distant the input signal is to the receptive field, the lower is its activity level; e.g., **Figure 3B**.



2.3. REINFORCEMENT LEARNING AND ERROR REWARD PROCESSING

The use of the rank-order coding algorithm provides an easy framework for reinforcement learning and error-based learning (Barto, 1995). For instance, the adaptation of the weights in Equation 3 can be modified simply with a variable $\alpha \in [0, 1]$ that can ponder Δw ; see Equation 4. If $\alpha = 0$, then the weights are not reinforced: $W_{t+1} = W_t$. If $\alpha = 1$, then the weights are reinforced in the direction of ΔW : $W_{t+1} = W_t + \alpha \Delta W$. In addition, conditional learning can be made simply by summing an external bias β to the neurons output X . By changing the amplitude of the neurons, we change also the rank-order to be learned and influence therefore the long-term the overall organization of the network; see Equation 5.

$$\Delta w \leftarrow \alpha \Delta w, \alpha \in [0, 1] \quad (4)$$

$$X \leftarrow X + \beta, \beta \in [-1, +1] \quad (5)$$

2.3.1. Cortical plasticity in PPC

For modeling the cortical plasticity in the PPC output maps, we implement an experience-driven plasticity mechanism. Observations done in rats show that during the learning of novel motor skills the synapses rapidly spread in the neocortex immediately as the animal learns a new task (Xu et al., 2009; Ziv and Ahissar, 2009). Rougier and Boniface proposed a dynamic learning rule in self-organizing maps to combine both the stability of the synapses' population to familiar inputs and the plasticity of the synapses' population to novel patterns (Rougier and Boniface, 2011). In order to model this feature in our PPC map, we redefine the coefficient α in Equation 5 and we rearrange the formula proposed by Rougier and Boniface:

$$\alpha = e^{1/\eta^2 / ||\max(X^{PPC}) - X_s^{PPC}||} \in [0, 1] \quad (6)$$

where η is the elasticity or plasticity parameter that we set to 1 and $\max(X^{PPC})$ is the upper bound of the neural activity, its maximal value, whereas $\max(X^{PPC})$ is the current maximum value within the neural population, with $\alpha = 0$ when $X_s^{PPC} = \max(X^{PPC})$. In this equation, the winner neuron learns the data according to its own distance to the data. If the winner neuron is close enough to it, it converges slowly to represent the data. At contrary, if the winner neuron is far from the data, it converges rapidly to it.

2.3.2. Error-reward function in ACC

For modeling ACC, we implement an error-reward function similar to Khamassi et al. (2011) and to Q-learning based algorithms. The neurons' value is updated afterwards only when an error occurs, then a inhibitory feedback error signal is sent to the winning neuron to diminish its activity X_{win} : $ACC(X_{win}) = -1$; the neurons equation X is updated as follows:

$$X_n^{PPC} = \sum_{m \in M} \frac{1}{rank(I_m)} w_{n,m} + ACC(X_n^{PPC}). \quad (7)$$

The neurons activity in ACC is cleared everytime the system responds correctly or provides a good answer. ACC can be seen then as a contextual working memory, a saliency buffer

extracted from the current context when errors occur inhibiting the wrong actions performed. Its activity may permit to establish an exploration-based type of learning by trial and errors and an attentional switch signal from automatic responses, in order to deal with the unexpected when a novel situation occurs.

2.4. PFC—SPATIO-TEMPORAL LEARNING IN A RECURRENT NETWORK

We can employ the rank-order coding for modeling spike-based recurrent neural network in which the amplitude values of the incoming input signals are replaced by its past spatio-temporal activity pattern. Although the rank-order coding algorithm has been used at first to model the fast processing of the feed-forward neurons in V1, its action has been demonstrated to replicate also the hebbian learning mechanism of Spike Timing-Dependent Plasticity (STDP) in cortical neurons (Bi and Poo, 1998; Abbott and Nelson, 2000; Izhikevich et al., 2004). For a population of N neurons, we arbitrarily choose to connect each neuron to a buffer of size $20 \times N$ so that they encode the rank code of the neurons amplitude value over the past 20 iterations. At each iteration, this buffer is shifted to accept the new values of the neurons.

$$X_n^{PFC} = \sum_{m \in M} \frac{1}{rank(buffer_m)} w_{n,m} + X_n^{PPC}. \quad (8)$$

Recurrent networks can generate novel patterns on the fly based on their previous activity pattern while, at each iteration, a winning neuron gets its links reinforced. Over time, the system regulates its own activity whereas coordinated dynamics can be observed. These behaviors can be used for anticipation and predictive control.

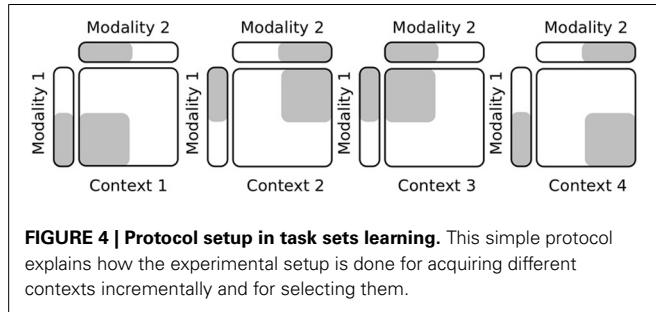
3. RESULTS

We propose to study the overall behavior of each neural system during the learning of task sets and the dynamics of the ensemble working together. The first three experiments are performed in a computer simulation only. They describe the behavior of the PPC maps working solely, working along the ACC system and working along the ACC and PFC systems for learning and selecting context-dependent task sets. Experiment 4 is performed on a robot arm. This experiment describes the acquisition and the learning of two different task set during the manipulation or not of a tool.

3.1. EXPERIMENT 1—PLASTICITY vs STABILITY IN LEARNING TASK SETS

In this first experiment, we test the capabilities of our network to learn incrementally novel contexts without forgetting the older ones, which corresponds to the so-called plasticity/stability dilemma of a memory system to retain the familiar inputs as well as to incorporate flexibly the novel ones. Our protocol follows the diagram in Figure 4 in which we show gradually four different contexts for two input modalities with vectors of ten indices. The input patterns are randomly selected from an area in the current context chosen randomly and for a period of time also variable. In this experiment, the PPC output map has 50 neurons that receive the activity of twenty gain-fields neurons, see Figure 2B.

We display in **Figure 5A** the raster plot of the PPC neurons' dynamics with distinct colors with respect to the context. Contexts are given gradually, one at a time, so that some neurons have to unlearn their previous cluster first in order to fit the new context. It is important to note that categorization is unsupervised and decided due to the experience-driven plasticity rule in Equation 6. In order to demonstrate the plasticity of the PPC network during the presentation of a new context, we present the context number four, plotted in magenta and never seen before, at $t = 11500$. Here, the new cluster is rapidly formed



and stable over time again to the cortical plasticity mechanism from Equation 6. The graph displays therefore not only the plasticity of the clusters in the PPC network but also their robustness.

This property is also shown in **Figure 5B** where the convergence rates of the PPC weights vary differently for each task. This result explains how the PPC self-organizes itself into different clusters that specialize flexibly with respect to the task. The ratio between stability and plasticity is shown in **Figure 5C** within the network with the histogram of the neuron's membership over a certain time interval. The stability of one neuron is computed as its probability distribution relative to each context. The higher values correspond to very stable neurons, which are set to one context only and do not deviate from it, whereas the lower values correspond to very flexible neurons that change frequently context from one to another.

The histogram shows two probability distributions within the system and therefore two behaviors. For the neurons corresponding to values near the strong peak at 1.0, their activity is very stable and strongly identified to one context. This shows that for one third of the neurons, the behavior of the neural population is very stable. At reverse, the power law curve centered on 0.0 shows the

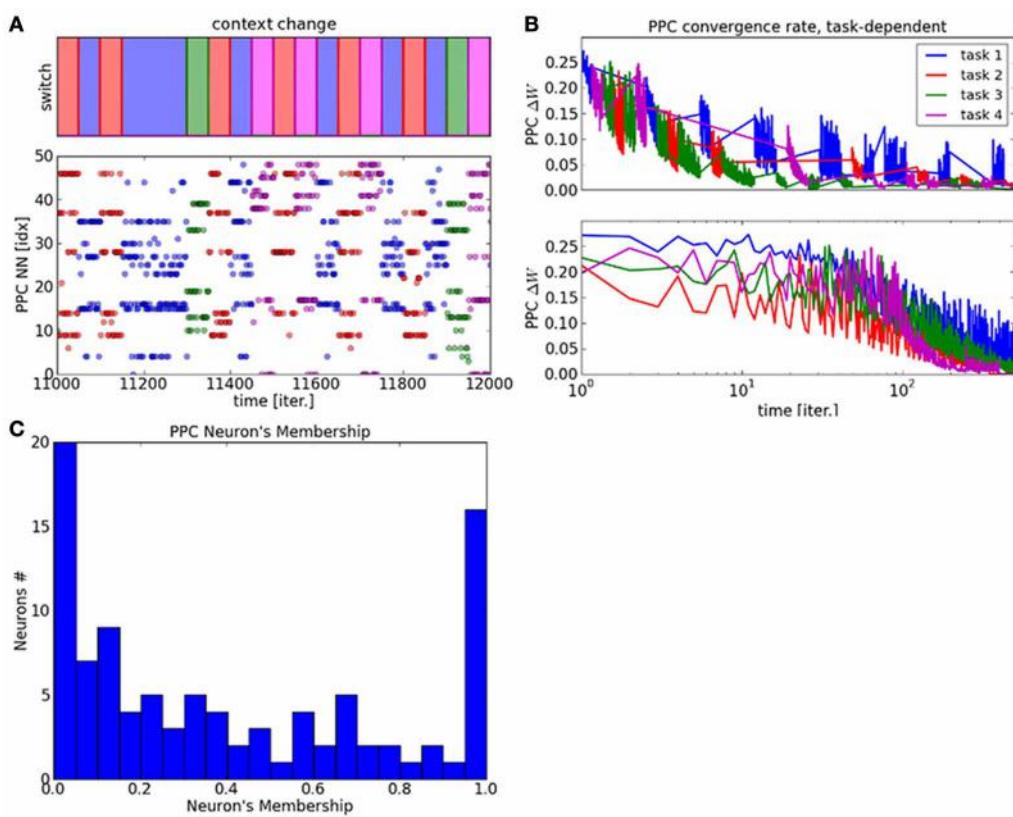


FIGURE 5 | Raster plot of the PPC output map and plasticity vs. stability within the map. **(A)** the graph displays the neural dynamics during task switch among four different contexts. **(B)** Convergence rate of the PPC network with respect to each task. **(C)** The degree of plasticity and stability within the PPC output map is represented as the probability distribution of the neurons membership to the cluster relative

to a context. This histogram shows two behaviors within the system. On the one hand, one third of the neurons present very stable dynamics with membership to one context only. On the other hand, two thirds of the neurons are part of different clusters and therefore to different contexts. The later neurons follow a power law distribution showing very plastic dynamics.

high variability of certain neurons, which are very dynamic for one third of the neural population.

We study now the neurons' activity during a task switch in **Figure 6**. In graph (A), the blue lines correspond to the neurons' dynamic belonging to the context before the switch and the red lines correspond to the neurons' dynamic belonging to the context after the switch. The activity level in each cluster is very salient for each context. The probability distribution of the neurons' dynamic, with respect to each context is plotted in **Figure 6B**. It shows a small overlap between the contexts before and after the switch.

3.2. EXPERIMENT 2—LEARNING TASK SETS WITH A REINFORCEMENT SIGNAL

In this second experiment, we reproduce a decision-making problem similar to those done in monkeys and humans with multiple choices and rewards (Churchland and Ditterich, 2012). The rules are not given in advance and the tasks switch randomly after a certain period of time with no regular pattern. The goal of the experiment is to catch the input-output correspondence pattern to stop making the error. The patterns are learned dynamically by reinforcement learning within each map and should ideally be done without interference from each other. The error signal indicates when an input-output association is erroneous with respect to a hidden policy, however, we make the note that it does not provide any hint about how to minimize the error. To understand how the whole system works, we focus our experiment on the PPC network with the ACC error processing system first, then with the PFC network. We choose to perform a two-choices experiment, with two output PPC maps initialized with random connections from the PPC map. The PPC network consists therefore of the gain-field architecture with the two output maps for modeling the two contexts. The two maps are then bidirectionally linked to the ACC system; the input signals for modality 1 and 2 are projected to the PPC input vectors of twenty units each; map1 has twelve output units and map2 has thirteen output units and project to ACC of dimension twenty-five units.

The hidden context we want the PPC maps to learn is to have output signals activated for specific interval range of the inputs

signals, namely, the first output map has to be activated when input neurons of indices below ten are activated, and reciprocally, the second output map has to be activated when input neurons of indices above ten are activated—this corresponds to the two first contexts in **Figure 4**. The error prediction signal is updated anytime a mistake has been done on the interval range to learn. As expressed in the previous section, the ACC error signal resets always its activity when the PPC maps start to behave correctly.

We analyze the performance of the PPC-ACC system in the following. We display in **Figures 7A,B** the raster plots of the PPC and ACC dynamics with respect to the context changes for different periods of time. The chart on the top displays the timing for context switch, the chart on the middle plots the ACC system working memory and the chart below plots the output of the PPC units. The **Figure 7A** is focusing on the beginning of the learning phase and the **Figure 7B** when the system has converged. We observe from these graphs that the units of the output maps self-organize very rapidly to avoid the error. ACC modulates negatively the PPC signals. We make the note that the error signal does not explicitly inhibit one map or the other but only the wrongly activated neuron of the map. As it can be observed, over time, each map specializes to its task. As a result, learning is not homogenous and depends also to the dimension of the context; that is, each map learns with a different convergence rate. ACC error rapidly reduces its overall activity for the learning of task1 with respect to map1, although the error persists for the learning of task2 with respect to map2 where some neurons still fires wrongly.

We propose to study the convergence of the two maps and the confidence level of the overall system for the two tasks. We define a confidence level index as the difference of amplitude between the most active neurons in map1 and map2. We plot its graph in **Figure 8** where the blue color corresponds to the confidence level for task1 with $v_{s_map1} - v_{s_map2}$ and the color red corresponds to the confidence level for task2 with $v_{s_map2} - v_{s_map1}$ during the learning phase. The dynamics reproduce similar trends from **Figure 7** where the confidence level constantly progresses till convergence to a stable performance rate, with a

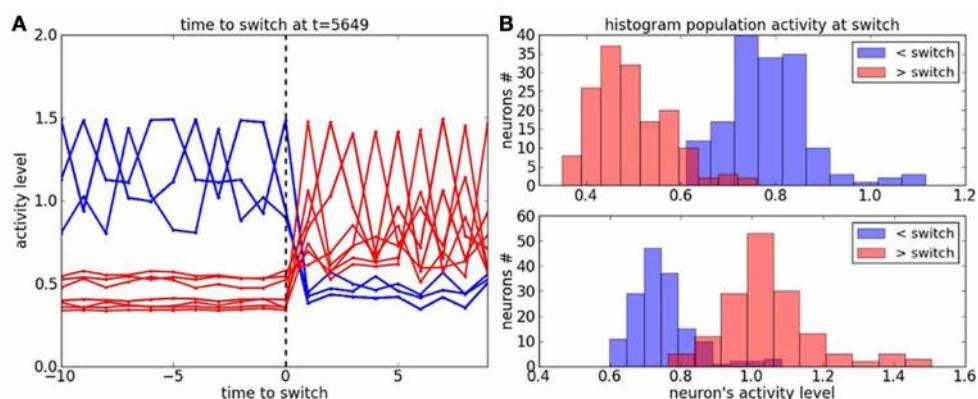


FIGURE 6 | Cluster dynamics at the time to switch. (A) Neural dynamics of the active clusters before and after the switch; resp. in blue and in red. **(B)** Histogram of the neural population at the time to switch with respect to the active clusters before and after the switch.

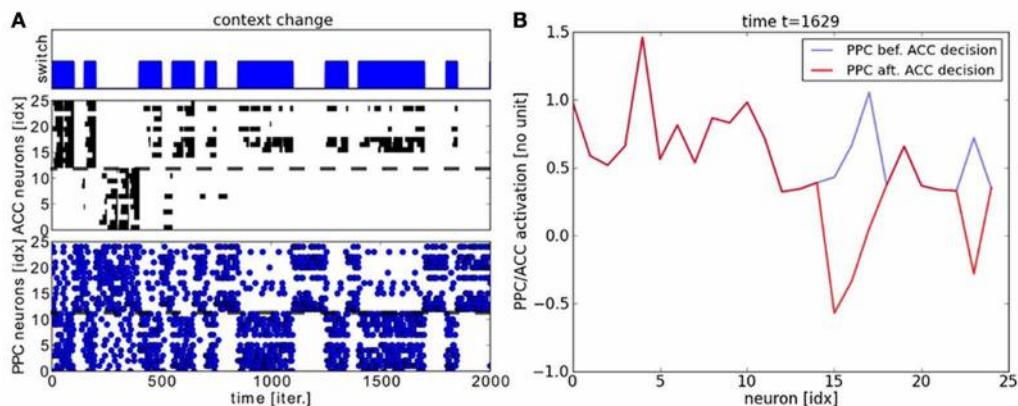


FIGURE 7 | Experiment on two-choices decision making and task switching. **(A)** Neural dynamics of PPC neurons and ACC error system during task switch. We plot in the chart in the top the temporal interval for each task. Below the, neural dynamics of the PPC maps and in the middle, its erroneous activity retranscribed in the ACC system. ACC works as a working

memory that keep tracks of the erroneous outputs, which is used during the learning stage. ACC is reset each time the PPC system gives a correct answer. Through reinforcement learning, the PPC maps converge gradually to the correct probability distribution. **(B)** Snapshot of the PPC maps in blue modulated negatively by ACC in red.

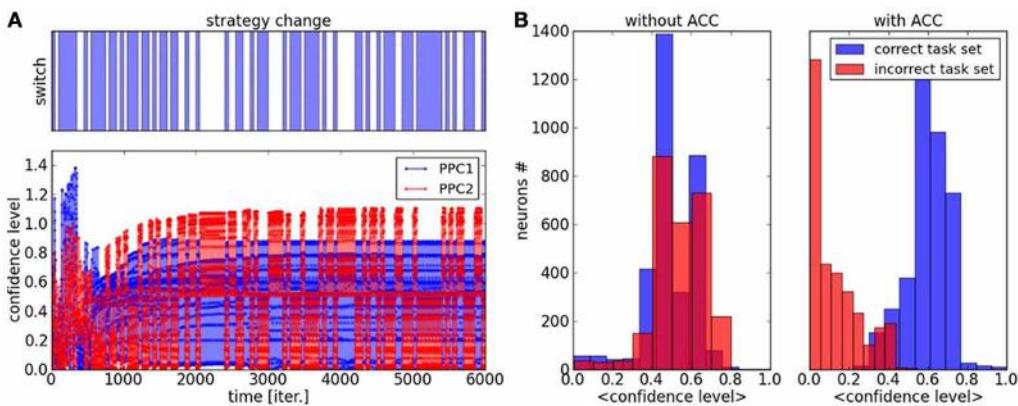


FIGURE 8 | Confidence Level of PPC maps during task switch, dynamics and histogram. **(A)** The confidence level is the difference between the amplitude of most activated neuron and the second one within each map. After one thousand iterations, the two maps rapidly specialize their dynamics to its associated task. This behavior is due to the ACC error-based learning.

(B) histogram of the probability distribution of the confidence level with and without ACC. With ACC, we observe a clear separation in two distributions, which correspond to a decrease of uncertainty with respect to the task. In comparison, the confidence level in an associative network without an error feedback gives a uniform distribution.

threshold around 0.4 above which a contextual state is recognized or not. Before 1000 iterations, the maps are very plastic so the confidence level fluctuates rapidly and continuously between different values but at the end of the learning phase, the maps are more static so the confidence level appears more discrete. This state is clearly observable from the histogram of the confidence level plotted on the right in **Figure 8B** for the case where the ACC error signal is injected to the associative network. The graph presents a probability distribution with two bell-shaped centred on 0.1 and 0.7, which corresponds to the cases of recognition or not of the task space. In comparison, the probability distribution for the associative learning without error-feedback is uniform, irrespective to the task; see **Figure 8B** in blue.

3.3. EXPERIMENT 3—ADAPTIVE LEARNING ON A TEMPORAL SEQUENCE BASED ON ERROR PREDICTION REWARD

We attempt to replicate now Harlow's experiments on adaptive learning, but, in comparison to the previous experiments, it is the temporal sequence of task sets that is taken into account for the reward. We employ our neural system in a cognitive experiment first to learn multiple spatio-temporal sequences and then to predict when a change of strategy has occurred based on the error or on the reward received. With respect to the previous section also, we add the PFC-like recurrent neural network to learn the temporal sequence from the PPC and ACC signals, see **Figure 1**.

The experiment is similar to the previous two-choices decision-making task, expect that the inputs follow now a temporal sequence within each map. When the inputs reach a particular

point in the sequence—, a point to switch,— we proceed to a random choice between one or the two trajectories. As in the previous section, the learning phase for the PPC rapidly converges to the specialization of the two maps thanks to the ACC error-learning processing. Meanwhile, the PFC learns the temporal organization of the PPC outputs based on their sequential order, **Figure 9A**. We do not give to the PFC any information about length, the number of patterns or the order of the sequence. Besides, each firing neuron reinforces its links with the current pre-synaptic neurons; see the raster plot in **Figure 9B**. After the learning phase, each PFC neuron has learned to predict some portion of the sequence based on the past and current PFC activity. Their saliency to the current sequence is retranscribed in their amplitude level. We plot the activity level of the neurons #10 and #14 respectively in black and red in the second chart. This graph shows that their activity level gradually increases for period intervals of at least ten iterations till their firing. The points to switch are also learned by the network and they are observable when the variance of the neurons' activity level becomes low, which is also seen when the confidence level goes under 0.4; which corresponds to the dashed black line in the first chart. For instance, we plot the dynamics of the PPC neurons and of the PFC neurons during such situation in **Figure 10A** at time $t = 1653$. The neural dynamics of each map display different patterns and therefore, different decisions. The PPC activates more the neurons of the first map (the neurons with indices below thirteen in blue) whereas the PFC activates more the neurons of the second map (the neurons with indices above thirteen in dashed red). This shows that the PFC is not a purely passive system driven by the current activity in PPC/ACC. Besides, it learns also to predict the future events based on its past activity. The PFC fuses the two systems in its dynamics, and this is why it generates here a noisy output distribution due to the conflicting signals. We plot in **Figure 10B** the influence of PPC on the PFC dynamics. In 60% of the cases, the two systems agree to predict the current dynamics. This corresponds

to the case of an automatic response when familiar dynamics are predicted. During conflicts, a prediction error is done by one of the two systems and in more cases the PPC dynamics, modulated by ACC, overwrite the values of the PFC units (blue bar). This situation occurs during a task switch for instance. At reverse, when PFC elicits its own values with respect to PPC (red bar), this situation occurs more when there is ambiguous sensory information that can be overpassed.

In order to understand better the decision-making process within the PFC map, we display in **Figures 11A,B** the temporal integration done dynamically at each iteration within the network. Temporal integration means the process of summing the weights in Equation 2 at each iteration with respect to the current order. If the sequence order is well recognized, then the neuron's value goes high very rapidly, otherwise its value remains to a low value. As we explained it in the previous paragraph, each neuron is sensitive to certain patterns in the current sequence based on the synaptic links within the recurrent network. This is translated in the graph by the integration of bigger values. The spatio-temporal sequences they correspond to are darkened proportionally to their activation level. The higher is the activation level integration during the integration period, the faster is the anticipation of the sequence. We present the cases for a unambiguous pattern in **Figure 11A** and for an ambiguous sequence activity in **Figure 11B**. The case for a salient sequence recognition in **Figure 11A** indicates that the current part of the sequence is well estimated by at least one neuron, the winning neuron, which predicts well the sequence over twenty steps in advance, see the chart below. In comparison, the dynamics in **Figure 11B** show a more uniform probability distribution. This situation arises when a bifurcation point is near in the sequence, it indicates that the system cannot predict correctly the next steps of the sequence.

Considering the decision-making process per se, there is not a strict competition between the neurons, however, each neuron

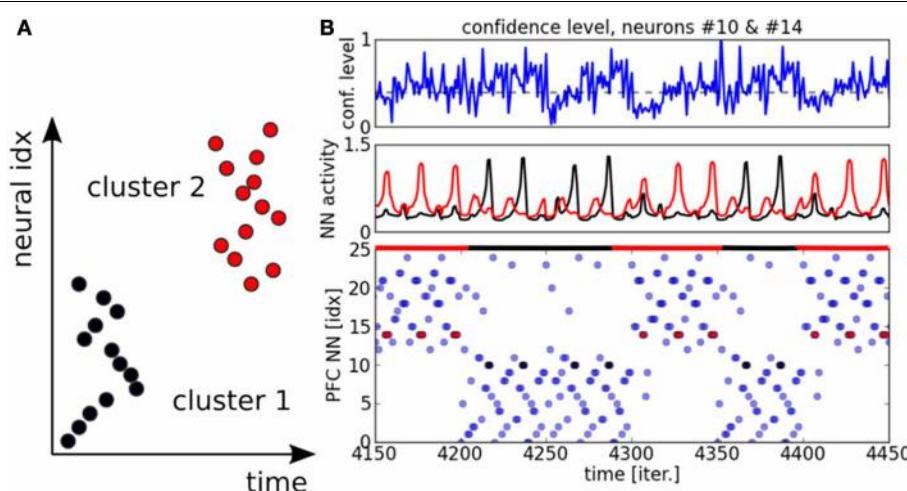
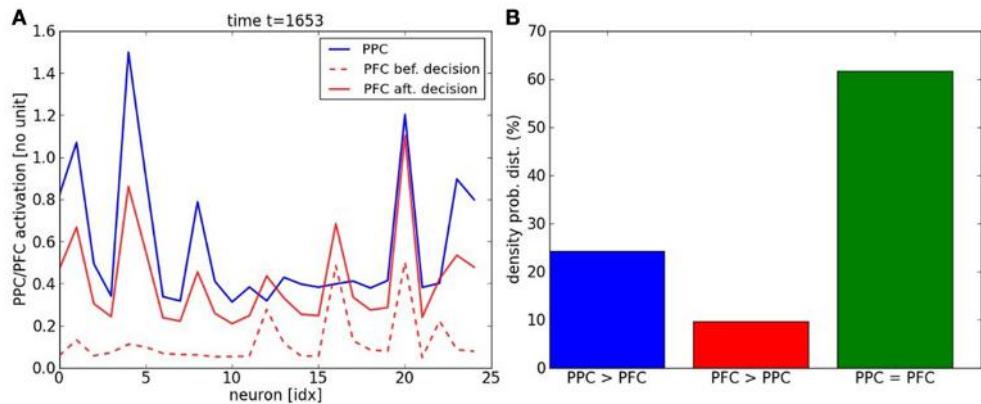
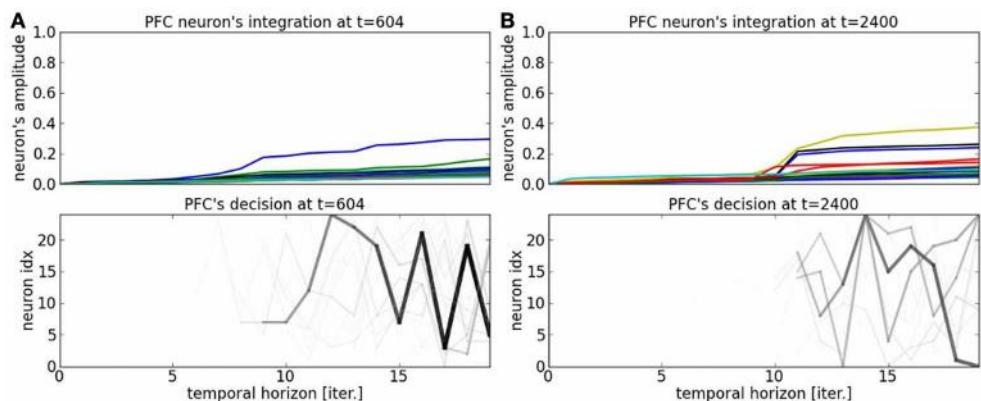


FIGURE 9 | Raster plot for PFC neurons. In (A), the PFC learns the particular temporal sequence from PPC outputs and it is sensitive to the temporal order of each unit in the sequence. In (B) on the top chart, the confidence level on the incoming signals from the PPC is plotted. The chart

in the middle displays the neural activity for two neurons from the two distinct clusters. The neuron #10 in black (resp. cluster #1) and the neuron #14 in red (resp. cluster #2). The raster plot of the whole system is plotted in the chart below.



values of PFC units, when PFC elicits its own values with respect to PPC and when both agree on the current predict. The PPC-PFC system works mostly in coherence from each other for 60% of the time (green bar) but in situations of conflict, the PPC overwrites twice the dynamics of the PFC network (blue bar) than the reverse (red bar).



sequence is detected, the farther the prediction of the trajectory. (B) At bifurcation points, the trajectories are fuzzier and several patterns are elicited.

promotes one spatio-temporal sequence and one probability distribution. Therefore, we have within the system 25 spatio-temporal trajectories embedded. Based on the current situation, some neurons will detect better one portion of the sequence than others and the probability distribution will be updated in consequence to chain the actions sequentially, whereas other portions will collapse. The decision-making looks therefore similar to a self-organization process.

At this point, no inhibitory system has been implemented directly in PFC that would avoid a conflict in the sequence order. Instead, the PFC integrates the PPC signals with the ACC error signals. The temporal sequences done in the PPC to avoid the errors at the next moves are learned little by little by reinforcement in the PFC. These sequences become strategies for error avoidance and explorative search. Over time, they learn the prediction of reward and the prediction of errors (Schultz et al., 1997; Schultz and Dickinson, 2000).

We perform some functional analysis on the PFC network in **Figure 12**. The connectivity circle in **Figure 12A** can permit to visualize the functional organization of the network at the neurons' level. We subdivide the PFC network into two sub-maps corresponding to the task dynamics in blue and red. We draw the strong intra-map connections between the neurons in the same color to their corresponding sub-maps as well as the strong inter-map connections between neurons of each map. Each neuron has a different connectivity in the network and the more it has connection the more it is central in the network. These neurons propagate information within and between the sub-maps, see **Figure 12B**. In complex systems terms, they are hub-like neurons from which different trajectories can be elicited. In decision-making, they are critical points for changing task. The density probability distribution plotted in **Figure 12C** shows that the maximum number of connections per neuron with strong synaptic weights reaches the number of four connections.

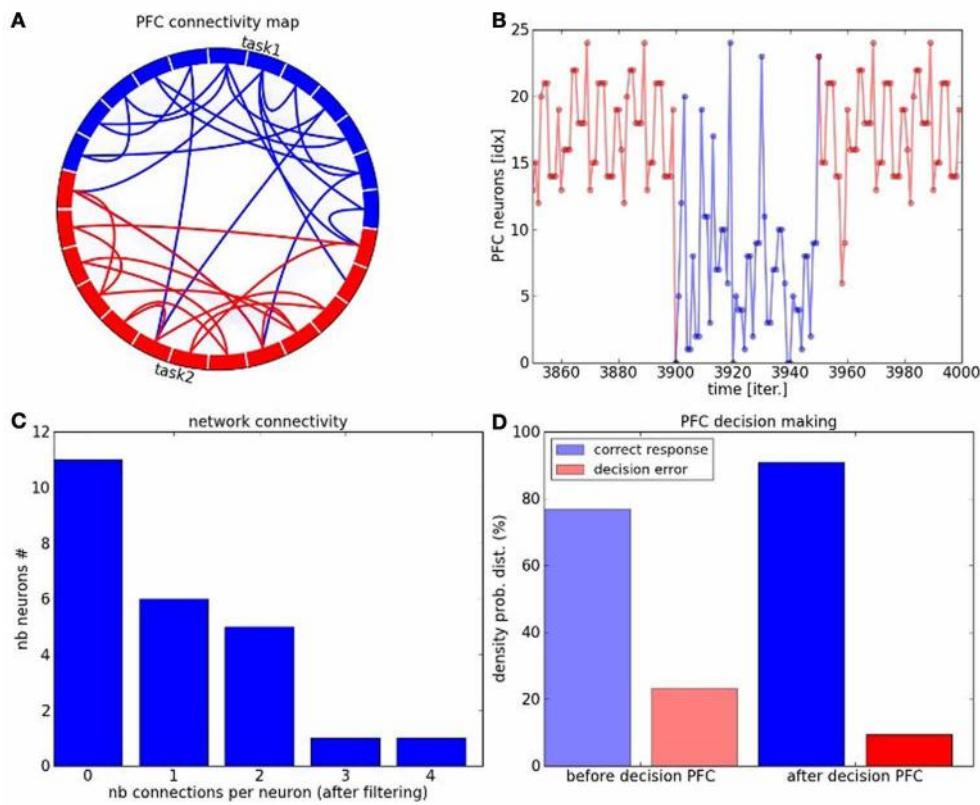


FIGURE 12 | PFC network analysis. **(A)** Connectivity circle for the neurons of the PFC map. In blue are displayed the neurons belonging to cluster 1 and in red are displayed the neurons belonging to cluster 2. The number of links within each cluster (intra-map connectivity) is higher than the number of links between them (inter-map connectivity). Moreover, the number of highly connected neurons is also weak. These characteristics replicate the ones of complex systems and of small-world networks in particular. **(B)** Task switch is

done through these hub-like neurons which can direct the trajectory from one or the other task. **(C)** The connectivity level per neurons within the network follows a logarithmic curve typical of complex networks, where the mostly connected neurons are also the fewer and the most critical with 4 distant connections. **(D)** The PFC network contributes to enhance the decision-making process in comparison to the PPC-ACC system due to the learning of the temporal sequence and to its better organization.

Their number drastically diminishes with respect to the number of connections and their trend follows a logarithmic curve. These characteristics correspond to the properties of small-world and scale-free networks.

In Figure 12D, we analyze the performance of the overall system when the PFC is added. The decision-making done in the PFC permits to decrease the error by a factor two: ten percent error in comparison to experiment 2. The prediction done in the recurrent map shows that the PFC is well organized to anticipate rewards and also task switch.

3.4. EXPERIMENT 4—ROBOTIC EXPERIMENT ON SENSORIMOTOR MAPPING AND ACTION SELECTION

We want to perform now a robotic experiment on action selection and decision making in the motor domain with a robotic arm of 6 degrees of freedom from the company Kinova; see Figure 13. We inspire ourselves on the one hand from Wolpert's experiments on structural learning and representation of uncertainty in motor learning (Wolpert and Flanagan, 2010; Orban and Wolpert, 2011) and on the other hand from Iriki's experiments on the spatial adaptation following active tool-use (Iriki et al., 1996; Maravita

and Iriki, 2004). Here, we attempt to learn different relations between states and motor commands when the robot controls its own arm alone and when it handles a tool. The question arises whether the robot will learn the structural affordances of the tool as a distinct representation or, instead, as part of its limb's representation (Cothros et al., 2006; Kluzik et al., 2008). Iriki et al. (1996) reported that bimodal-cell visual receptive fields (vRFs) show spatial adaptation following active tool-use, but not passive holding. The spatial estimation of its own body limits—that is, its body image—is different depending on the attention to the tool. The goal is therefore to estimate properly the current situation on which the robot is, which means handling a stick or not, actively or passively.

In our framework, we expect that the errors of spatial estimation on the end-point can be gradually learned and that sensorimotor mapping will change with respect to the tasks the robot has to perform (Wolpert and Flanagan, 2010; Orban and Wolpert, 2011). Figures 13A,B display the arm robot when it holds a salient toy and when it handles a stick with the toy at its end-point. In this experiment, a fixed camera is mapping the x-y coordinates of the salient points (i.e., the toy) while the robot

moves its arm around its elbow; we make the note that we circumscribe the problem to two modalities only in order to control just one articulation with respect to the Y axis in the camera.

In the previous experiments, we did not exploit specifically the properties of the gain-field neurons for mapping sensorimotor

transformation. Here instead, we use the gain-field mechanism to combine the visuomotor information into the PPC system for the two contexts. With respect to the task, the PPC output maps will learn the specific amplitude of the gain-field neurons corresponding to the specific visuomotor relationships (Holmes et al., 2007).

For instance, we plot in **Figures 14A–D** the activity level of four different gain-field neurons relative to the motor angle θ_0 of the robot arm. The blue dots represent the situation when it weaves the hand in front of the camera and the red dots represent the situation when it is handling the tool. As the gain-field neurons learn the specific relationship between certain values of the XY coordinates of the end-point effector and the motor angle θ_0 , this value is modulated when the robot arm uses the stick; see resp. **Figures 14A–D**. The visuo-motor translation in the XY plane when the robot is handling the tool produces a gain modulation that decreases or increases the neurons' activity level.

Hence, the visuomotor coordination changes instantaneously the GF neurons' activity level relative to the current task set and the PPC is dynamically driven by the input activity (not displayed). The neural activity in the PFC map, instead, can evolve autonomously and independently with respect to the input activity, even if the PPC dynamics are presented for a short exposure; this behavior is displayed in the raster plot in **Figure 15A**.

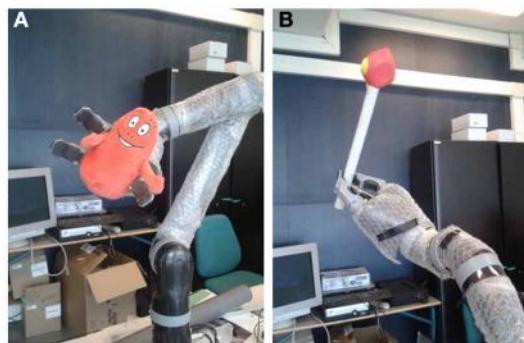


FIGURE 13 | Robot arm Kinova for task-set selection. The two task-sets correspond to **(A)** the situation when it is moving its hand alone with the red target on its hand and **(B)** the situation when it is moving the stick on its hand with the red target on the tip of the tool.

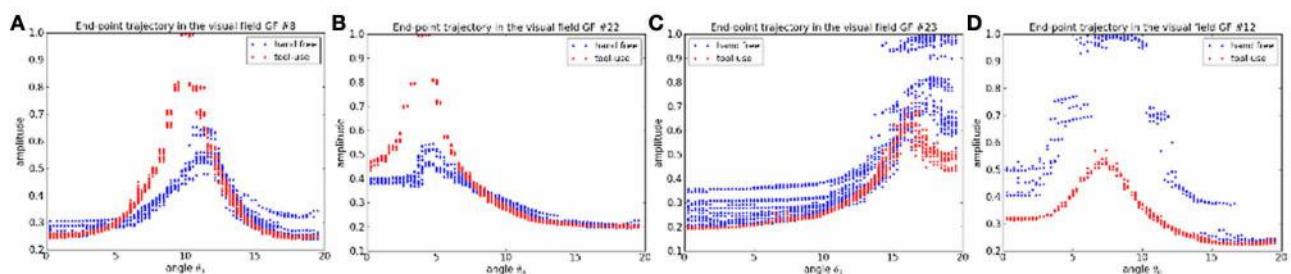


FIGURE 14 | Dynamics of the gain-field neurons relative to the task. (A–D) In blue, the robot moves its hand freely. In red, the robot is handling the tool. Depending on what the GF neurons have learned, their peak level will diminish or increase when changing the task (i.e., using a tool).

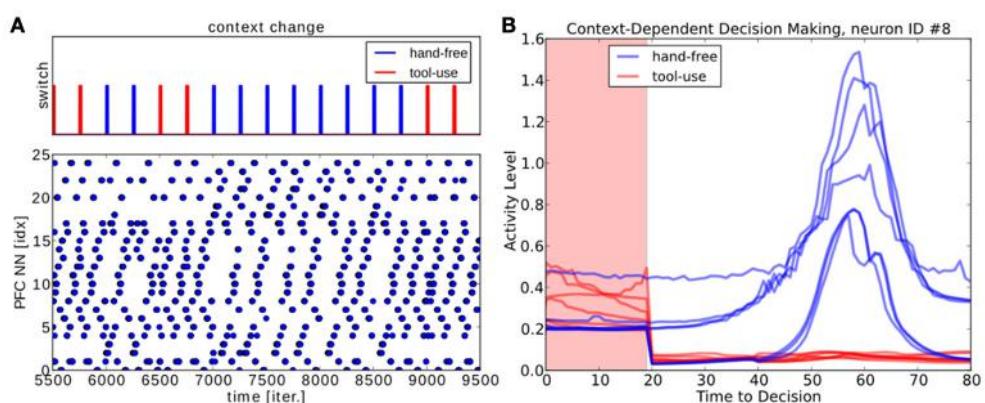


FIGURE 15 | PFC Attention decision during contextual change, hand-free or tool-use. **(A)** we expose to the PFC dynamics some incomplete patterns for a short period of time of 20 iterations, every 500 iterations. The PFC is

capable to switch to the reconstruct back the missing part of the spatio-temporal sequence; in blue for hand-free and in red for tol-use. **(B)** Neural activity for one neuron when one of the two contexts is set.

When we expose the PFC neurons to the PPC dynamics for a small period of time—20 iterations every 500 iterations (the segments on the top chart),—the network is able to reconstruct dynamically the rest of the ongoing sequence; see **Figure 15B**. For instance, the neuron #8 is selective to the particular context of hand-free (blue lines). The contextual information is maintained as a stable pattern of the neural activity in the working memory and the contexts are accessible and available for influencing the ongoing processing. As a recurrent network, the PFC behaves similarly to a working memory. It embeds the two different strategies depending on the context, even in presence of incomplete inputs and can select to attend or not to the tool.

4. DISCUSSION

The ability to learn the structure of actions and to select on the fly the proper one given the current task is one great leap in infants cognition. During development, infants learn to differentiate their motor behaviors relative to various contexts by exploring and identifying the correct structures of causes and effects that they can perform by trial and errors. This behavior corresponds to an intrinsic motivation, a mechanism that is argued to drive cognitive development. Besides, Karen Adolph emphasizes the idea of “learning-to-learn” in motor development, an expression akin to Harlow that appears in line with the one of intrinsic motivation. She proposes that two learning mechanisms embody this concept during the development of the motor system—, respectively an associative memory and a category-based memory,— and that the combination of these two learning systems is involved in this capacity of learning-to-learn. Braun et al. (2010) foster a similar concept and suggest that motor categorization requires 1) a critic for learning the structure, i.e., an error-based system, and 2) a learning system that will learn the conditional relationships between the incoming variables; which means, the parameters of the task. They argue that once these parameters are found, it is easier to transfer knowledge from one initial task to many others. All-in-all, we believe that these different concepts on structural learning are important to scaffold motor development and to have intrinsic motivation in one system. Thus the question arises what are the neural mechanisms involved in structural learning and in flexible behaviors?

To investigate this question, we have modeled an architecture that attempts to replicate the functional organization of the fronto-parietal structures, namely, a sensorimotor mapping system, an error-processing system and a reward predictor (Platt and Glimcher, 1999; Westendorff et al., 2010). The fronto-parietal cortices are involved in activities related to observations of alternatives and to action planning, and the anterior cingulate cortex is a part of this decision-making network. Each of these neural systems contribute to one functional part of it. The ACC system is processing the error-negativity reward to the PPC maps for specialization and to the PFC network for reward prediction. The PPC network organizes the sensorimotor mapping for different tasks whereas the PFC learns the spatio-temporal patterns during the act.

In particular, the PPC is organized around the mechanism of gain-modulation where the gain-fields neurons combine the

sensory inputs from each other. We suggest that the mechanism of gain-modulation can implement the idea of structural learning in motor tasks proposed by Braun and Wolpert (Braun et al., 2009, 2010). In their framework, the gain-field neurons can be seen as basis functions and as the parameters of the learning problem. It is interesting to note that Braun and al. make a parallel with the bayesian framework, which has been also proposed to describe the gain-field mechanism. For instance, Deneve explains the computational capabilities of gain-fields in the context of the bayesian framework to efficiently represent the joint distribution of a set of random variables (Denève and Pouget, 2004).

Parallelly, we used three specific intrinsic mechanisms for enhancing structural learning: the rank-order coding algorithm, the cortical plasticity and an error-based reward. For instance, the rank-order coding algorithm was used to emulate efficiently the so-called spike timing-dependent plasticity to learn spatio-temporal sequences in a recurrent network (Bi and Poo, 1998; Abbott and Nelson, 2000). The PFC system exploits their properties for self-organizing itself by learning the sequences of each task as well as the switch points. PFC neurons learn specific trajectories and at each iteration, a competition process is at work to promote the new steps of the ongoing sequence. Besides, cortical plasticity was modeled in PPC maps with an activity-dependent learning mechanism that promotes the rapid learning of novel (experienced-based) tasks and the stabilization of the old ones. An advantageous side-effect of this mechanism is that PPC neurons become context-dependent, which is a behavior observed also in the reaching neurons of the parieto-motor system, the so-called mirror neurons (Gallese et al., 1996; Brozovic et al., 2007). The results found on cortical plasticity are in line with observations on the rapid adaptation of the body image and of the motor control. Wolpert observed that the motor system incorporates a slow learning mechanism along a fast one for the rapid formation of task sets (Wolpert and Flanagan, 2010). The cortical plasticity is also influenced by an error-based system in ACC that reshape the PPC dynamics with respect to the task. The negative reward permits to inhibit the wrong dynamics but not to elicit the correct ones. Those ones are gradually found by trial and errors, which replicate an exploration process.

We believe that these different mechanisms are important for incremental learning and intrinsic motivation. However, many gaps remain. For instance, a truly adaptive system should show more flexibility during familiar situations than during unfamiliar ones. Retranscribed from Adolph and Joh (2005), a key to flexibility is (1) to refrain from forming automatic responses and (2) to identify the critical features that allow online problem solving to occur. This ability is still missing in current robots. In the context of problem solving in tool-use, Fagard and O'Regan emphasizes the similar difficulty for infants to use a stick for reaching a toy. They also observe that below a certain age, attention is limited to one object only as they just cannot “hold in mind” the main goal in order to perform one subgoal (Fagard et al., 2012; Rat-Fischer et al., 2012). Above this period, however, Fagard and O'Regan observe an abrupt transition in their behaviors when they became capable to relate two actions at a time, to plan consecutive actions and to use recursion. They hypothesize that after

16 months, infants are able to enlarge their focus of attention to two objects simultaneously and to “bufferize” the main goal. We make a parallel with the works of Koechlin and colleagues Koechlin et al. (2003); Collins and Koechlin (2012) who attribute a monitoring role to the frontal cortex for maintaining the working memory relative to the current tasks and for prospecting the different action sequences or episodic

memories (Koechlin and Summerfield, 2007), which will be our next steps.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the French ANR projects INTERACT and NEUROBOT, and the Centre National de la Recherche Scientifique (CNRS).

REFERENCES

- Abbott, L., and Nelson, S. (2000). Synaptic plasticity: taming the beast. *Nat. Neurosci.* 3, 1178–1182. doi: 10.1038/81453
- Adolph, K. (2008). Learning to move. *Curr. Dir. Psychol. Sci.* 17, 213–218. doi: 10.1111/j.1467-8721.2008.00577.x
- Adolph, K., and Joh, A. (2005). “Multiple learning mechanisms in the development of action,” in *Paper presented to the Conference on Motor Development and Learning*. Vol. 33, eds J. Lockman, J. Reiser, and C. A. Nelson, (Murcia).
- Adolph, K., and Joh, A. (2009). *Multiple Learning Mechanisms in the Development of Action*. New York, NY: Oxford University Press.
- Andersen, R. (1997). Multimodal integration for the representation of space in the posterior parietal cortex. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 353, 1421–1428. doi: 10.1098/rstb.1997.0128
- Andersen, R., and Cui, H. (2009). Intention, action planning, and decision making in parietal-frontal circuits. *Neuron* 63, 568–583. doi: 10.1016/j.neuron.2009.08.028
- Barto, A. (1995). “Adaptive critics and the basal ganglia,” in *Models of Information Processing in the Basal Ganglia* eds J. Houk, J. Davis, and D. Beiser (Cambridge, MA: MIT Press), 215–232.
- Baumann, M., Fluet, M. C., and Scherberger, H. (2009). Context-specific grasp movement representation in the macaque anterior intraparietal area. *J. Neurosci.* 29, 6436–6448. doi: 10.1523/JNEUROSCI.5479-08.2009
- Bi, G., and Poo, M. (1998). Activity-induced synaptic modifications in hippocampal culture, dependence of spike timing, synaptic strength and cell type. *J. Neurosci.* 18, 10464–10472.
- Botvinick, M., Braver, T., Barch, D., Carter, C., and Cohen, J. (2001). Conflict monitoring and cognitive control. *Psychol. Rev.* 108, 624–652. doi: 10.1037/0033-295X.108.3.624
- Braun, D., Aertsen, A., Wolpert, D., and Mehring, C. (2009). Motor task variation induces structural learning. *Curr. Biol.* 19, 352–357. doi: 10.1016/j.cub.2009.01.036
- Braun, D., Mehring, C., and Wolpert, D. (2010). Structure learning in action. *Behav. Brain Res.* 206, 157–165. doi: 10.1016/j.bbr.2009.08.031
- Brozovic, M., Gail, A., and Andersen, R. (2007). Gain mechanisms for contextually guided visuomotor transformations. *J. Neurosci.* 27, 10588–10596. doi: 10.1523/JNEUROSCI.2685-07.2007
- Churchland, A., and Ditterich, J. (2012). New advances in understanding decisions among multiple alternatives. *Curr. Opin. Neurobiol.* 22, 920–926. doi: 10.1016/j.conb.2012.04.009
- Collins, A., and Koechlin, E. (2012). Reasoning, learning, and creativity: frontal lobe function and human decision-making. *PLoS Biol.* 10:e1001293. doi: 10.1371/journal.pbio.1001293
- Cothros, N., Wong, J., and Gribble, P. (2006). Are there distinct neural representations of object and limb dynamics? *Exp. Brain Res.* 173, 689–697. doi: 10.1007/s00221-006-0411-0
- Denève, S., and Pouget, A., (2004). Bayesian multisensory integration and cross-modal spatial links. *J. Neurophysiol.* 98, 249–258. doi: 10.1016/j.jnphysparis.2004.03.011
- Diamond, A. (1985). Development of the ability to use recall to guide action, as indicated by infants’ performance on a-not-b. *Child Dev.* 74, 24–40.
- Diamond, A. (1990). Rate of maturation of the hippocampus and the developmental progression of children’s performance on the delayed non-matching to sample and visual paired comparison tasks. *Ann. N.Y. Acad. Sci.* 608, 394–426; discussion: 426–433. doi: 10.1111/j.1749-6632.1990.tb48904.x
- Fagard, J., Rat-Fischer, L., and O’Regan, J. (2012). Comment le bb accorde-t-il la notion d’outil? *Enfance* 64, 73–84. doi: 10.4074/S0013754512001085
- Fluet, M., Baumann, M., and Scherberger, H. (2010). Context-specific grasp movement representation in the macaque ventral premotor cortex. *J. Neurosci.* 30, 15175–1518. doi: 10.1523/JNEUROSCI.3343-10.2010
- Fuster, J. (2001). The prefrontal cortex: Time is of the essence. *Neuron* 30, 319–333. doi: 10.1016/S0896-6273(01)00285-9
- Gallesi, V., Fadiga, L., Fogassi, L., and Rizzolatti, G. (1996). Action recognition in the premotor cortex. *Brain* 119, 593–609. doi: 10.1093/brain/119.2.593
- Guerin, F., Kruger, N., and Kraft, D. (2013). A survey of the ontogeny of tool use: from sensorimotor experience to planning. *Aut. Ment. Dev. IEEE Trans.* 5, 18–45. doi: 10.1109/TAMD.2012.2209879
- Harlow, H. (1949). The formation of learning sets. *Psychol. Rev.* 56, 51–65. doi: 10.1037/h0062474
- Holmes, N., Sanabria, D., Calvert, G., and Spence, C. (2007). Tool-use: capturing multisensory spatial attention or extending multisensory peripersonal space? *Cortex* 43, 469–489. doi: 10.1016/S0010-9452(08)70471-4
- Holroyd, C., and Coles, M. (2002). The neural basis of human error processing: reinforcement learning, dopamine, and the error-related negativity. *Psychol. Rev.* 109, 679–709. doi: 10.1037/0033-295X.109.4.679
- Holtmaat, A., and Svoboda, K. (2009). Experience-dependent structural synaptic plasticity in the mammalian brain. *Nat. Rev. Neurosci.* 10, 647–665. doi: 10.1038/nrn2699
- Iriki, A., Tanaka, M., and Iwamura, Y. (1996). Coding of modified body schema during tool use by macaque postcentral neurones. *Neuroreport* 7, 2325–2330. doi: 10.1097/00001756-199610020-00010
- Izhikevich, E., Gally, A., and Edelman, G. (2004). Spike-timing dynamics of neuronal groups. *Cereb. Cortex* 14, 933–944. doi: 10.1093/cercor/bhh053
- Johnson, M. (2012). Executive function and developmental disorders: the flip side of the coin. *Trends Cogn. Sci.* 16, 454–457. doi: 10.1016/j.tics.2012.07.001
- Kaplan, F., and Oudeyer, P.-Y. (2007). In search of the neural circuits of intrinsic motivation. *Front. Neurosci.* 1, 225–236. doi: 10.3389/neuro.01.1.017.2007
- Khamassi, M., Lallâïe, S., Enel, P., Procyk, E., and Dominey, P. (2011). Robot cognitive control with a neurophysiologically inspired reinforcement learning model. *Front. Neurorobot.* 5, 1–14. doi: 10.3389/fnbot.2011.00001
- Kluzik, J., Diedrichsen, J., Shadmehr, R., and Bastian, A. (2008). Reach adaptation: what determines whether we learn an internal model of the tool or adapt the model of our arm? *J. Neurophysiol.* 100, 1455–1464. doi: 10.1152/jn.90334.2008
- Koechlin, E., Ody, C., and Kouneiher, F. (2003). The architecture of cognitive control in the human prefrontal cortex. *Science* 302, 1181–1185. doi: 10.1126/science.1088545
- Koechlin, E., and Summerfield, C. (2007). An information theoretical approach to prefrontal executive function. *Trends Cogn. Sci.* 11, 229–235. doi: 10.1016/j.tics.2007.04.005
- Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biol. Cybern.* 43, 59–69. doi: 10.1007/BF00337288
- Lockman, J. (2000). A perception-action perspective on tool use development. *Child Dev.* 71, 137–144. doi: 10.1111/1467-8624.00127
- Maravita, A., and Iriki, A. (2004). Tools for the body (schema). *Trends Cogn. Sci.* 8, 79–86. doi: 10.1016/j.tics.2003.12.008
- Orban, G., and Wolpert, D. (2011). Representations of uncertainty in sensorimotor control. *Curr. Opin. Neurobiol.* 21, 1–7. doi: 10.1016/j.conb.2011.05.026
- Pitti, A., Blanchard, A., Cardinaux, M., and Gaussier, P. (2012). “Gain-field modulation mechanism in multimodal networks for spatial perception,” *12th IEEE-RAS International Conference on Humanoid Robots Nov.29-Dec.1, 2012*. Business Innovation Center Osaka, 297–302.
- Pitti, A., and Kuniyoshi, Y. (2011). “Modeling the cholinergic

- innervation in the infant cortico-hippocampal system and its contribution to early memory development and attention," *Proceedings of the International Joint Conference on Neural Networks (IJCNN11)*, (San Jose, CA), 1409–1416. doi: 10.1109/IJCNN.2011.6033389
- Platt, M., and Glimcher, P. (1999). Neural correlates of decision variables in parietal cortex. *Nature* 400, 233–238. doi: 10.1038/22268
- Pouget, A., and Snyder, L. (2000). Computational approaches to sensorimotor transformations. *Nat. Neurosci.* 3, 1192–1198. doi: 10.1038/81469
- Rat-Fischer, L., O'Regan, J., and Fagard, J. (2012). The emergence of tool use during the second year of life. *J. Exp. Child Psychol.* 113, 440–446. doi: 10.1016/j.jecp.2012.06.001
- Rougier, N., and Boniface, Y. (2011). Dynamic self-organising map. *Neurocomputing* 74, 1840–1847. doi: 10.1016/j.neucom.2010.06.034
- Salinas, E., and Thier, P. (2000). Gain modulation: a major computational principle of the central nervous system. *Neuron* 27, 15–21. doi: 10.1016/S0896-6273(00)00004-0
- Schöner, G., and Dineva, E. (2007). Dynamic instabilities as mechanisms for emergence. *Dev. Sci.* 10, 69–74. doi: 10.1111/j.1467-7687.2007.00566.x
- Schultz, W., Dayan, P., and Montague, P. (1997). A neural substrate of prediction and reward. *Annu. Rev. Neurosci.* 27, 1593–1599.
- Schultz, W., and Dickinson, A. (2000). Neuronal coding of prediction errors. *Annu. Rev. Neurosci.* 23, 473–500. doi: 10.1146/annurev.neuro.23.1.473
- Singh, S., Lewis, R., Barto, A., and Sorg, J. (2010). Intrinsically motivated reinforcement learning: an evolutionary perspective. *IEEE Trans. Autonom. Mental Dev.* 2, 70–82. doi: 10.1109/TAMD.2010.2051031
- Smith, L., Thelen, E., Titzer, R., and McLin, D. (1999). Knowing in the context of acting: The task dynamics of the a-not-b error. *Psychol. Rev.* 106, 235–260. doi: 10.1037/0033-295X.106.2.235
- Stricanne, B., Andersen, R., and Mazzon, P. (1996). Eye-centered, head-centered, and intermediate coding of remembered sound locations in area lip. *J. Neurophysiol.* 76, 2071–2076.
- Tenenbaum, J., Kemp, C., Griffiths, T., and Goodman, N. (2011). How to grow a mind: statistics, structure, and abstraction. *Science* 331, 1279–1285. doi: 10.1126/science.1192788
- Thorpe, S., Delorme, A., and Van Rullen, R. (2001). Spike-based strategies for rapid processing. *Neural Netw.* 14, 715–725. doi: 10.1016/S0893-6080(01)00083-1
- Vaesen, K. (2012). The cognitive bases of human tool use. *Behav. Brain Sci.* 35, 203–262. doi: 10.1017/S0140525X11001452
- Van Rullen, R., Gautrais, J., Delorme, A., and Thorpe, S. (1998). Face processing using one spike per neurone. *Biosystems* 48, 229–239. doi: 10.1016/S0303-2647(98)00070-7
- Van Rullen, R., and Thorpe, S. (2002). Surfing a spike wave down the ventral stream. *Vision Res.* 42, 2593–2615. doi: 10.1016/S0042-6989(02)00298-5
- Warren, J., and Harlow, H. (1952). Learned discrimination performance by monkeys after prolonged postoperative recovery from large cortical lesions. *J. Comp. Physiol. Psychol.* 45, 119–126. doi: 10.1037/h0055350
- Westendorff, S., Klaes, C., and Gail, A. (2010). The cortical timeline for deciding on reach motor goals. *J. Neurosci.* 30, 5426–5436. doi: 10.1523/JNEUROSCI.4628-09.2010
- White, O., and Diedrichsen, J. (2013). Flexible switching of feedback control mechanisms allows for learning of different task dynamics. *PLoS ONE* 8:e54771. doi: 10.1371/journal.pone.0054771
- Wolpert, D., Diedrichsen, J., and Flanagan, J. (2011). Principles of sensorimotor learning. *Nat. Rev. Neurosci.* 12, 739–751. doi: 10.1038/nrn3112
- Wolpert, D., and Flanagan, J. (2010). Motor learning. *Curr. Biol.* 20, R467–R472. doi: 10.1016/j.cub.2010.04.035
- Xu, T., Yu, X., Perlak, A., Tobin, W., Zweig, J., Tennant, K., et al. (2009). Rapid formation and selective stabilization of synapses for enduring motor memories. *Nature* 462, 915–919. doi: 10.1038/nature08389
- Yokoyama, C., Tsukada, H., Watanabe, Y., and Onoe, H. (2005). A dynamic shift of neural network activity before and after learning-set formation. *Cereb. Cortex* 15, 796–801. doi: 10.1093/cercor/bhh180
- Ziv, N., and Ahissar, E. (2009). New tricks and old spines. *Nature* 462, 859–861. doi: 10.1038/462859a

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 20 June 2013; accepted: 01 October 2013; published online: 22 October 2013.

Citation: Pitti A, Braud R, Mahé S, Quoy M and Gaussier P (2013) Neural model for learning-to-learn of novel task sets in the motor domain. *Front. Psychol.* 4:771. doi: 10.3389/fpsyg.2013.00771

This article was submitted to Cognitive Science, a section of the journal *Frontiers in Psychology*.

Copyright © 2013 Pitti, Braud, Mahé, Quoy and Gaussier. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Curiosity driven reinforcement learning for motion planning on humanoids

Mikhail Frank^{1,2,3*}, Jürgen Leitner^{1,2,3}, Marijn Stollenga^{1,2,3}, Alexander Förster^{1,2,3} and Jürgen Schmidhuber^{1,2,3}

¹ Dalle Molle Institute for Artificial Intelligence, Lugano, Switzerland

² Facoltà di Scienze Informatiche, Università della Svizzera Italiana, Lugano, Switzerland

³ Dipartimento Tecnologie Innovative, Scuola Universitaria Professionale della Svizzera Italiana, Manno, Switzerland

Edited by:

Gianluca Baldassarre, Italian National Research Council, Italy

Reviewed by:

Anthony F. Morse, University of Skövde, Sweden

Hsin Chen, National Tsing-Hua University, Taiwan

Alberto Finzi, Università di Napoli Federico II, Italy

***Correspondence:**

Mikhail Frank, Dalle Molle Institute for Artificial Intelligence, Galleria 2, CH-6928 Manno-Lugano, Switzerland

e-mail: kail@idsia.ch

Most previous work on *artificial curiosity* (AC) and *intrinsic motivation* focuses on basic concepts and theory. Experimental results are generally limited to toy scenarios, such as navigation in a simulated maze, or control of a simple mechanical system with one or two degrees of freedom. To study AC in a more realistic setting, we *embody* a curious agent in the complex iCub humanoid robot. Our novel reinforcement learning (RL) framework consists of a state-of-the-art, low-level, reactive control layer, which controls the iCub while respecting constraints, and a high-level curious agent, which explores the iCub's state-action space through information gain maximization, learning a world model from experience, controlling the actual iCub hardware in real-time. To the best of our knowledge, this is the first ever embodied, curious agent for real-time motion planning on a humanoid. We demonstrate that it can learn compact Markov models to represent large regions of the iCub's configuration space, and that the iCub explores *intelligently*, showing *interest* in its physical constraints as well as in objects it finds in its environment.

Keywords: **artificial curiosity, intrinsic motivation, reinforcement learning, humanoid, iCub, embodied AI**

1. INTRODUCTION

Reinforcement Learning (RL) (Barto et al., 1983; Sutton and Barto, 1998; Kaelbling et al., 1996) allows an *agent* in an *environment* to learn a *policy* to maximize some sort of *reward*. Rather than optimizing the policy directly, many RL algorithms instead learn a value function, defined as expected future discounted cumulative reward. Much of early RL research focused on discrete states and actions instead of continuous ones dealt with by function approximation and feature-based representations.

An RL agents needs to *explore* its environment. Undirected exploration methods (Barto et al., 1983), rely on randomly selected actions, and do not differentiate between already explored regions and others. Contrastingly, directed exploration methods can focus the agent's efforts on novel regions. They include the classic and often effective *optimistic initialization*, go-to the least-visited state, and go-to the least recently visited state.

1.1. ARTIFICIAL CURIOSITY (AC)

Artificial Curiosity (AC) refers to directed exploration driven by a world model-dependent value function designed to direct the agent toward regions where it can learn something. The first implementation (Schmidhuber, 1991b) was based on an *intrinsic reward* inversely proportional to the predictability of the environment. A subsequent AC paper (Schmidhuber, 1991a) emphasized that the reward should actually be based on the *learning progress*, as the previous agent was motivated to fixate on inherently unpredictable regions of the environment. Subsequently, a probabilistic AC version (Storck et al., 1995) used the well known

Kullback-Leibler (KL) divergence (Lindley, 1956; Fedorov, 1972) to define non-stationary, intrinsic rewards reflecting the changes of a probabilistic model of the environment after new experiences. Itti and Baldi (2005) called this measure *Bayesian Surprise* and demonstrated experimentally that it explains certain patterns of human visual attention better than previous approaches.

Over the past decade, robot-oriented applications of curiosity research have emerged in the closely related fields of Autonomous Mental Development (AMD) (Weng et al., 2001) and Developmental Robotics (Lungarella et al., 2003). Inspired by child psychology studies of Piaget (Piaget and Cook, 1952), they seek to learn a strong base of useful skills, which might be combined to solve some externally posed task, or built upon to learn more complex skills.

Curiosity-driven RL for developmental learning (Schmidhuber, 2006) encourages the learning of appropriate skills. Skill learning can be made more explicit by identifying learned skills (Barto et al., 2004) within the option framework (Sutton et al., 1999). A very general skill learning setting is assumed by the PowerPlay framework, where skills actually correspond to arbitrary computational problem solvers (Schmidhuber, 2013; Srivastava et al., 2013).

Luciw et al. (2011) built a curious planner with a high-dimensional sensory space. It learns to perceive its world and predict the consequences of its actions, and continually plans ahead with its imperfect but optimistic model. Mugan and Kuipers developed QLAP (Mugan and Kuipers, 2012) to build predictive models on a low-level visuomotor space. Curiosity-Driven Modular Incremental Slow Feature

Analysis (Kompella et al., 2012) provides an intrinsic reward for an agent's progress toward learning new spatiotemporal abstractions of its high-dimensional raw pixel input streams. Learned abstractions become option-specific feature sets that enable skill learning.

1.2. DEVELOPMENTAL ROBOTICS

Developmental Robotics (Lungarella et al., 2003) seeks to enable robots to learn to do things in a general and adaptive way, by trial-and-error, and it is thus closely related to AMD and the work on curiosity-driven RL, described in the previous section. However, developmental robotic implementations have been few.

What was possibly the first AC-like implementation to run on hardware (Huang and Weng, 2002) rotated the head of the SAIL robot back and forth. The agent/controller was rewarded based on reconstruction error between its improving internal perceptual model and its high-dimensional sensory input.

AC based on learning progress was first applied to a physical system to explore a playroom using a Sony AIBO robotic dog. The system (Oudeyer et al., 2007) selects from a variety of pre-built behaviors, rather than performing any kind of low-level control. It also relies on a remarkably high degree of random action selection, 30%, and only optimizes the immediate (next-step) expected reward, instead of the more general delayed reward.

Model-based RL with curiosity-driven exploration has been implemented on a Katana manipulator (Ngo et al., 2012), such that the agent learns to build a tower, without explicitly rewarding any kind of stacking. The implementation does use pre-programmed *pick and place* motion primitives, as well as a set of specialized pre-designed features on the images from an overhead camera.

A curiosity-driven modular reinforcement learner has recently been applied to surface classification (Pape et al., 2012), using a robotic finger equipped with an advanced tactile sensor on the fingertip. The system was able to differentiate distinct tactile events, while simultaneously learning behaviors (how to move the finger to cause different kinds of physical interactions between the sensor and the surface) to generate the events.

The so-called hierarchical curiosity loops architecture (Gordon and Ahissar, 2011) has recently enabled a 1-DOF LEGO Mindstorms arm to learn simple reaching (Gordon and Ahissar, 2012).

Curiosity implementations in developmental robotics have sometimes used high dimensional sensory spaces, but each one, in its own way, greatly simplified the action spaces of the robots by using pre-programmed high-level motion primitives, discretizing motor control commands, or just using very, very simple robots. We are unaware of any AC (or other intrinsic motivation) implementation, which is capable of learning in, and taking advantage of a complex robot's high-dimensional configuration space.

Some methods learn internal models, such as hand-eye motor maps (Nori et al., 2007), inverse kinematic mappings (D'Souza et al., 2001), and operational space control laws (Peters and Schaal, 2008), but these are not curiosity-driven. Moreover, they lack the generality and robustness of full-blown path planning algorithms (Latombe et al., 1996; LaValle, 1998; Li and Shie, 2007; Perez et al., 2011).

1.3. THE PATH PLANNING PROBLEM

The *Path Planning Problem* is to find motions that pursue goals while deliberately avoiding arbitrary non-linear constraints, usually obstacles. The ability to solve the path planning problem in practice is absolutely critical to the eventual goal of deploying complex/humanoid robots in unstructured environments. The recent textbook, "Planning Algorithms" (LaValle, 2006), offers many interesting approaches to planning motions for complex manipulators. These are expensive algorithms, which search the configuration space to generate trajectories that often require post-processing. Thus robots, controlled by algorithmic planners, are typically very deliberate and slow, first "thinking," often for quite some time, then executing a motion, which would be simple and intuitive for humans.

1.4. REACTIVE CONTROL

In the 1980s, a control strategy emerged, which was completely different from the established *plan first, act later* paradigm. The idea was to use potential fields (Khatib, 1986; Kim and Khosla, 1992), and/or dynamical systems (Schoner and Dose, 1992; Iossifidis and Schoner, 2004, 2006), and/or the sensor signals directly (Brooks, 1991) to generate control commands fast, without searching the configuration space. Control is based on some kind of local gradient, which is evaluated at the robot's current configuration. As a result, sensors and actuators are tightly coupled in a fast, light weight action/observation loop, allowing a robot to react quickly and smoothly to changing circumstances. Nevertheless, reactive controllers are shortsighted and prone to getting stuck in local minima/maxima, making them relatively bad path planners.

1.5. A CURIOUS CONFLUENCE

In this paper, we introduce a curiosity-driven reinforcement learner for the iCub humanoid robot (Metta et al., 2008), which autonomously learns a powerful, reusable solver of motion planning problems from experience controlling the actual, physical robot.

The application of RL to the path planning problem (or more precisely the process of embodying the agent at a sufficiently low level of control) has allowed us to incorporate two approaches, planning and reactive control, which for the most part have been treated separately by roboticists until now. The integrated system benefits from both approaches while avoiding their most problematic drawbacks, and we believe it to be an important step toward realizing a practical, feasible, developmental approach to real, non-trivial robotics problems. Furthermore, the system is novel in the following ways:

1. In contrast to previous implementations of artificial curiosity and/or intrinsic motivation in the context of developmental robotics, our system learns to control many degrees of freedom (DOFs) of a complex robot.
2. Planning algorithms typically generate reference trajectories, which must then be passed to a controller. Our RL system, on the other hand, learns control commands directly, while still yielding a resolution complete planner. This greatly simplifies many practical issues that arise from tracking a reference

- trajectory and results in a lighter, faster action/observation loop.
3. Rather than relying on reactive control to generate entire motions, we only use it to implement actions. Thus the completeness of the planner is preserved, although its robustness is improved by the added capacity of each action react to unforeseen and/or changing constraints.

2. MATERIAL AND METHODS

In order to build a developmental learning system capable of exploiting the iCub's high DOF configuration space, we begin by looking at the path planning literature, where there exist two classes of algorithms, capable of generating high dimensional reference trajectories. Single query algorithms, such as Rapidly Exploring Random Trees (RRT) (LaValle, 1998; Perez et al., 2011), interpolate two points in configuration space, without reusing knowledge from one query to the next. Multiple query algorithms on the other hand, such as Probabilistic Road Maps (PRM) (Latombe et al., 1996; Sun et al., 2005), store a compressed representation of the configuration space and satisfy queries by operating on that data structure, rather than searching the high DOF configuration space directly. In the case of PRM, the configuration space is represented by a graph, which can even be grown incrementally (Li and Shie, 2007). PRM's compact, incrementally expandable representation of known motions makes it a likely antecedent to or template for a development learning system, but there are several problems, which are all related to *separation* between planning and control.

To build up a PRM planner, one must first sample the configuration space to obtain a set of vertices for the graph. The samples are then interpolated by trajectories, which form the set of edges that connect the vertices. The feasibility of each sample (vertex) and trajectory (edge) must be preemptively verified, typically by forward kinematics and collision detection computations, which collectively amount to a computationally expensive pre-processing step. The configuration of the robot *must* remain on the verified network of samples and trajectories at all times, or there may be unwanted collisions. This implies that all the trajectories in the graph must also be controllable, which is in general difficult to verify in simulation for complex robots, such as the iCub, which exhibit non-linear dynamics (due to friction and deformation) and are thus very difficult to model faithfully. If these problems can be surmounted, then a PRM planner can be constructed, however, the configuration of the robot's workspace must be static, because moving anything therein may affect the feasibility of the graph edges.

All of these problems can be avoided by *embodiment* of the planner and giving the system the capacity to *react*. If there were a low-level control system, which could enforce all necessary constraints (to keep the robot safe and operational) in real time, then the planner could simply try things out, without the need to exhaustively and preemptively verify the feasibility of each potential movement. In this case, reference trajectories would become unnecessary, and the planner could simply store, recall, and issue control commands directly. Lastly, and perhaps most importantly, with the capacity to react in real time, there would be no need to require a static workspace.

This new *embodied planner* would differ from its antecedent PRM planner in several important ways. There would be no need to require that the configuration of the robot be *on* any of the graph edges. In fact the graph would no longer represent a network of distinct trajectories, but rather the *topology* of the continuous configuration space. Each edge would no longer represent a particular trajectory, but rather a more general kind of *action* that implements something like *try to go to that region of the configuration space*. Such actions would be available not when the true robot configuration is *on* a graph vertex, but rather when it is *near* that vertex. The actions may or may not succeed depending on the particular initial configuration of the robot when the action was initiated as well as the configuration of the workspace, which must not necessarily be static.

Allowing the planner to control the hardware directly offers considerable benefits, but it also requires a more complex representation of the configuration space than the *plan first, act later* paradigm did. Whereas the PRM planner made do with a simple graph, representing a network of trajectories, the embodied version seems to require a probabilistic model, which can cope with actions that may have a number of different outcomes. In light of this requirement, the embodied planner begins to look like a Markov Decision Process (MDP), and in order to exploit such a planner, the state transition probabilities, which govern the MDP, must first be learned. However, this presents a problem in that experiments (trying out actions) are very expensive when run on robotic hardware, which is bound to real time, as opposed to simulations, which can be run faster than real time, or parallelized, or both. Therefore, an efficient exploration method is absolutely critical, which motivates our use of curiosity-driven RL.

2.1. ACTION IMPLEMENTATION

We have put considerable energy into developing the low-level control system described above, the Modular Behavioral Environment (MoBeE; **Figures 1** and **2**) (Frank et al., 2012), the details of which are beyond the scope of this paper. In this section, we describe MoBeE only insofar as to define the notion of *action* as it pertains to our RL system.

MoBeE controls the robot constantly, at a high frequency, according to the following second order dynamical system:

$$\mathbf{M}\ddot{\mathbf{q}}(t) + \mathbf{C}\dot{\mathbf{q}}(t) + \mathbf{K}(\mathbf{q}(t) - \mathbf{q}^*) = \sum f_i(t) \quad (1)$$

The vector function $\mathbf{q}(t) \in \mathbb{R}^n$ is the robot configuration, and the matrices \mathbf{M} , \mathbf{C} , and \mathbf{K} contain mass, damping, and spring constants, respectively. The position vector \mathbf{q}^* is an attractor, and constraints on the system are implemented by forcing it via $f_i(t)$, which provides automatic avoidance of kinematic infeasibilities having to do with joint limits, cable lengths, and collisions.

An action, for the purposes of RL, means setting the attractor \mathbf{q}^* to some desired configuration. When such an action is taken, $\mathbf{q}(t)$ begins to move toward \mathbf{q}^* . The action terminates either when the dynamical system settles or when a timeout occurs. The action

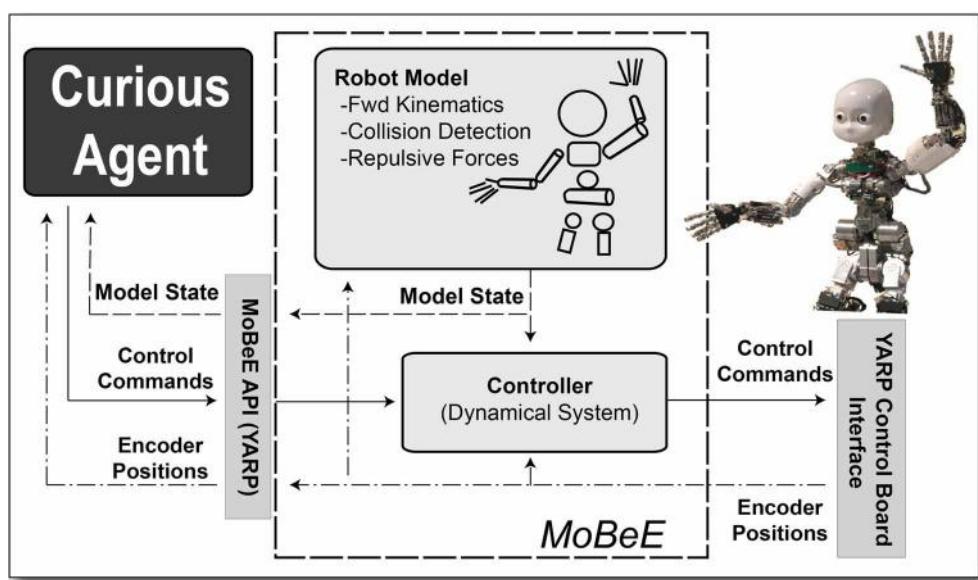


FIGURE 1 | MoBeE and the iCub. MoBeE (left) prevents the iCub humanoid robot (right) from colliding with the table. Semi-transparent geometries represent force fields, and when these

collide with one another (shown in red), they generate repulsive, constraint forces, which in this case push the hands away from the table surface.

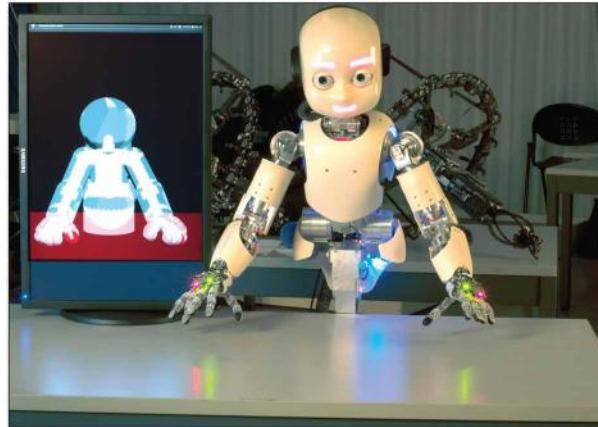


FIGURE 2 | The modular behavioral environment (MoBeE) architecture. MoBeE implements low-level control and enforces all necessary constraints to keep the robot safe and operational in real time, such that the curious RL agent (left) is able to experiment with arbitrary control commands. A kinematic/geometric model of the iCub humanoid robot (top) is driven by streaming motor encoder positions from the hardware (right). The model computes fictitious constraint forces, which repel the robot from collisions, joint limits, and other infeasibilities. These forces, $f_i(t)$ in Equation (1), are passed to the controller (middle), which computes the attractor dynamics that governs the actual movement of the robot.

may or may not settle on q^* , depending on what constraint forces, $f_i(t)$ are encountered during the transient response.

2.2. STATE-ACTION SPACE

The true configuration of the robot at any time t can be any real valued $q \in \mathbb{R}^n$, however, in order to define a tractable RL problem,

we discretize the configuration space (Figure 3) by selecting m samples, $Q = \{q_j | j = 1 \dots m\} \subset \mathbb{R}^n$. The sample set Q defines a set of states¹ $S = \{s_j | j = 1 \dots m\}$, such that $\bigcup_{j=1}^m s_j = \mathbb{R}^n$. Each state, $s_j \in S$, is the Voronoi region associated with the corresponding sample, $q_j \in Q$. That is to say, each sample, $q_j \in \mathbb{R}^n$, defines a state, $s_j \subset \mathbb{R}^n$, where every point, $q \in s_j$, is closer² to q_j than to any other point $q \in Q$. The states in our Markov model are the sets, $s \in S$, not the points, $q \in Q$, and to say that the robot is in some particular state, s , at some particular time, t , means that the real valued configuration of the robot, $q(t) \in s$.

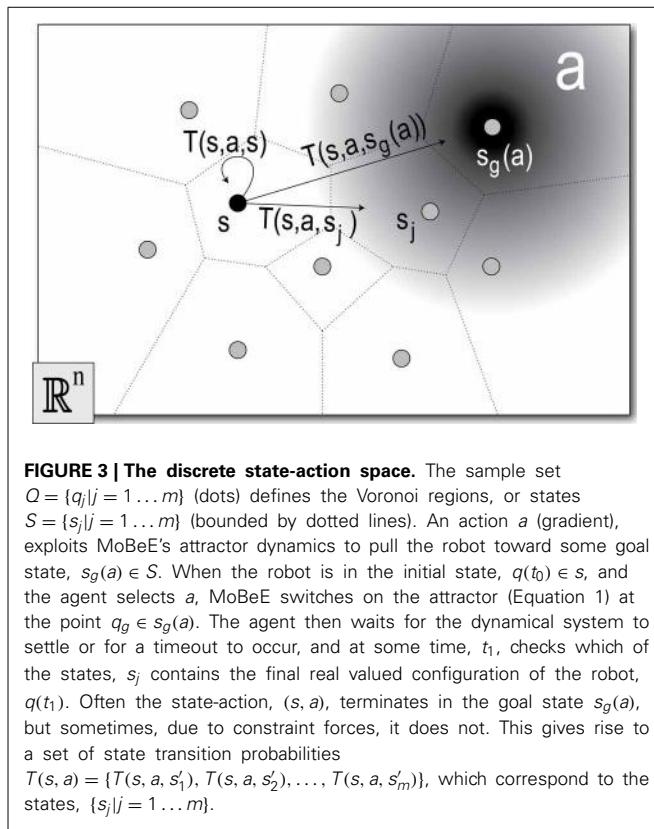
An action is defined by setting MoBeE's attractor, $q^* = q_g$ (Equation 1), where $q_g \in Q$ is the sample in some goal state $s_g(a)$. When an action is tried, the robot moves according to the transient response, $q(t)$, of the dynamical system, which eventually settles at $q(t \rightarrow \infty) = q_\infty$. However, depending on the constraint forces encountered, it may be that $q_\infty \in s_g(a)$ or not.

2.2.1. Connecting states with actions

An action, a , intends to move the robot to some goal state $s_g(a)$, a waypoint along the path that will eventually be generated by the reinforcement learner. But which states should be connected to which other states? In order that our Markov model develops into an effective path planner, we want to connect each

¹Generally, throughout this formalism we use uppercase letters to denote sets and lowercase letters to denote points. However, we have made an exception for the states, $s_j \in S$, which themselves comprise sets of robot configurations, $s_j \subset \mathbb{R}^n$. Although this is somewhat abusive from a set theoretic standpoint, it allows us to be consistent with the standard RL notation later in the paper.

²The distance metric employed is the Euclidean norm, in this case: $\sqrt{(q - q_j) \cdot (q - q_j)}$.



state to its k nearest neighbors³ in a way that makes sense with respect to the dimensionality of the configuration space, n . To this end, we choose $k = 2^n$, as an n -dimensional hypercube has 2^n vertices.

With each state, s , is associated a set of actions, $A(s)$, which intend to move the robot from s to each of k nearby goal states, $A(s) = \{a_g | g = 1 \dots k\}$, and the set of all possible actions, A , can therefore be expressed as the union of the action sets belonging to each state, $A = \bigcup_{s=1}^m A(s)$.

This notion of connecting neighboring states makes intuitive sense given the problem domain at hand and the resulting Markov model resembles the Roadmap graph used by the PRM planner (Latombe et al., 1996). Although the action set, A , is quite large ($|A| = |S|$), each state only has access to the actions, $A(s)$, which lead to its k nearest neighbors ($|A(s)| = k$). Therefore, the number of state-actions remains linear in the number of states. We advise the reader that wherever the standard state-action notation, (s, a) , is used, it is implied that $a \in A(s)$.

2.2.2. Modeling transition probabilities

Although each action *intends* to move the robot to some particular goal state, in principle they can terminate in *any* state in the set $\{s_j | j = 1 \dots m\}$. Therefore, we must learn state transition probabilities to represent the connectivity of the configuration

³Again, the distance metric employed is the Euclidean norm, in this case: $\sqrt{(q_g - q_i) \cdot (q_g - q_i)}$.

space. A straightforward way of doing this would be to define a probability distribution over all possible outcomes s_j for each state-action (s, a) :

$$T(q_\infty \in s_j | s, a) = \begin{cases} p(q_\infty \in s_1 | s, a) \\ p(q_\infty \in s_2 | s, a) \\ \vdots \\ p(q_\infty \in s_m | s, a) \end{cases} \quad (2)$$

To build up the distributions, $T(q_\infty \in s_j | s, a)$, we would simply initialize all probabilities to zero and then count the occurrences of observed transitions to the various states, s_j , resulting from the various state-actions (s, a) . We would, however, find this approach to be relatively wasteful, because much of the state-action space is deterministic. In practice, we find that there are only three kinds of distributions that come out of applying RL algorithms to our Markov model. A state-action, (s, a) , can terminate deterministically in the goal state $s_g(a)$ (Equation 3), it can terminate deterministically in some other state $s_j \neq s_g(a)$ (Equation 4), or it can be truly non-deterministic (Equation 5), although the non-zero components of T are always relatively few compared to the number of states in the model.

$$p(q_\infty \in s_j | s, a) = \begin{cases} 1 & \text{if } s_j = s_g(a) \\ 0 & \text{if } s_j \neq s_g(a) \end{cases} \quad (3)$$

$$p(q_\infty \in s_j | s, a) = \begin{cases} 1 & \text{if } s_j = s^* \neq s_g(a) \\ 0 & \text{if } s_j \neq s^* \end{cases} \quad (4)$$

$$p(q_\infty \in s_j | s, a) = \begin{cases} > 0 & \text{if } s_j \in S_1 \\ = 0 & \text{if } s_j \in S_0 \end{cases} \quad \left| \begin{array}{l} S_0 \cup S_1 = S, |S_0| \gg |S_1| \end{array} \right. \quad (5)$$

This is intuitive upon reflection. Much of the configuration space is not affected by constraints, and actions always complete as planned. Sometimes constraints are encountered, such as joint limits and cable length infeasibilities, which deflect the trajectory in a predictable manner. Only when the agent encounters changing constraints, typically non-static objects in the robot's operational space, do we see a variety of outcomes for a particular state-action. However, even in this case, the possible outcomes, s' , are a relatively small number of states, which are usually in the neighborhood of the initial state, s . We have never constructed an experiment, using this framework, in which a particular state-action, (s, a) , yields more than a handful of possible outcome states, s' .

We can and have used distributions of the form shown in Equation (2) to model the outcomes of state-actions in our RL framework. However, we have found a better way to represent the distribution, which is more parsimonious, and facilitates a better AC signal.

2.3. ARTIFICIAL CURIOSITY

What is interesting? For us humans, *interestingness* seems closely related to the rate of our learning progress (Schmidhuber, 2006). If we try doing something, and we rapidly get better at doing it, we are often interested. Contrastingly, if we find a task trivially

easy, or impossibly difficult, we do not enjoy a high rate of learning progress, and are often bored. We model this phenomenon using the information theoretic notion of *information gain*, or KL divergence.

2.3.1. KL divergence

KL Divergence, D_{KL} is defined as follows, where P_j and T_j are the scalar components of the discrete probability distributions P and T , respectively.

$$D_{KL}(P||T) = \sum_j \ln\left(\frac{P_j}{T_j}\right) P_j \quad (6)$$

For our purposes, T represents the estimated state transition probability distribution (Equation 2) for a particular state-action, (s, a) , after the agent has accumulated some amount of experience. Once the agent tries (s, a) again, an s' is observed, and the state transition probability distribution for (s, a) is updated. This new distribution, P , is a better estimate of the state transition probabilities for (s, a) , as it is based on more data.

By computing $D_{KL}(P||T)$, we can measure how much our Markov model improved by trying the state-action, (s, a) , and we can use this *information gain* to reward our curious agent. Thus, the agent is motivated to improve its model of the state-action space, and it will gravitate toward regions thereof, where learning is progressing quickly.

There is, however, a problem. The KL divergence is not defined if there exist components of P or T , which are equal to zero. This is somewhat inconvenient in light of the fact that for our application, most of the components of most of the distributions, T (Equation 2), are actually zero. We must therefore initialize P and T cleverly.

Perhaps the most obvious solution would be to initialize T with a uniform distribution, before trying some action for the first time. After observing the outcome of the selected action, P would be defined and $D_{KL}(P||T)$ computed, yielding the *interestingness* of the action taken.

Some examples of this kind of initialization are given in Equations (7–10)⁴. Clearly the approach solves the numerical problem with the zeros, but it means that initially, *every* action the agent tries will be equally interesting. Moreover, *how interesting* those first actions are, $|D_{KL}(P||T)|$, depends on the size of the state space.

$$D_{KL}(\{1, 2, 1\} || \{1, 1, 1\}) = 0.0589 \quad (7)$$

$$D_{KL}(\{2, 1, 1\} || \{1, 1, 1\}) = 0.0589 \quad (8)$$

$$D_{KL}(\{1, 1, 2, 1, 1\} || \{1, 1, 1, 1, 1\}) = 0.0487 \quad (9)$$

$$D_{KL}(\{1, 1, 1, 2, 1, 1, 1\} || \{1, 1, 1, 1, 1, 1, 1\}) = 0.0398 \quad (10)$$

The first two examples, Equations (7), (8), show that regardless of the outcome, all actions generate the same numerical interestingness the first time they are tried. While not a problem in

⁴We have intentionally not normalized P and T , to show how they are generated by counting observations of $q_\infty \in s_j$. In order to actually compute $D_{KL}(P||T)$, P and T must first be normalized.

Algorithm 1: Observe($s, a, s', T(s, a), R(s, a)$)

```

begin
  if there is no bin,  $T_{s'}(s, a)$ , in  $T(s, a)$  to count occurrences
  of  $s'$  then
    append a bin,  $T_{s'}(s, a)$  to  $T(s, a)$ 
     $T_{s'}(s, a) \leftarrow 1$ 
  end
   $P \leftarrow T(s, a)$ 
   $P_{s'} \leftarrow P_{s'} + 1$ 
   $R(s, a) \leftarrow D_{KL}(P||T(s, a))$ 
   $T(s, a) \leftarrow P$ 
end

```

theory, in practice this means our robot will need many tries to gather enough information to differentiate the boring, deterministic states from the interesting, non-deterministic ones. Since our actions are designed to take the agent to a goal state, $s_g(a)$, it would be intuitive if observing a transition to $s_g(a)$ were less interesting than observing one to some other state. This would drastically speed up the learning process.

The second two examples, Equations (9), (10) show that the *interestingness* of that first try decreases in larger state spaces, or alternatively, small state spaces are numerically more *interesting* than large ones. This is not a problem if there is only one learner operating in a single state-action space. However, in the case of a multi-agent system, say one learner per body part, it would be convenient if the intrinsic rewards gotten by the different agents were numerically comparable to one another, regardless of the relative sizes of those learners' state-action spaces.

In summary, we have two potential problems with KL Divergence as a reward signal:

1. Slowness of initial learning
2. Sensitivity to the cardinality of the distributions

Nevertheless, in many ways, KL Divergence captures exactly what we would like our curious agent to focus on. It turns out we can address both of these problems by representing T with an array of variable size, and initializing the distribution optimistically with respect to the expected behavior of the action (s, a) .

2.3.2. Dynamic state transition distributions

By compressing the distributions T and P , i.e., not explicitly representing any bins that contain a zero, we can compute the KL divergence between only their non-zero components. The process begins with T and P having no bins at all. However, they grow in cardinality as follows: Every time we observe a novel s' as the result of trying a state-action (s, a) , we append a new bin to the distribution $T(s, a)$, and initialize it with a 1, and copy it to yield $P(s, a)$. Then, since we just observed (s, a) result in s' , we increment the corresponding bin in $P(s, a)$, and compute $KL(P||T)$. This process is formalized in **Algorithm 1**.

The optimistic initialization is straightforward. Initially, the distribution $T(s, a)$ is empty. Then we *observe* (**Algorithm 1**) that (s, a) fails, leaving the agent in the initial state, s . The KL divergence between the trivial distributions $\{1\}$ and $\{2\}$ is 0, and

therefore, so is the reward, $R(s, a)$. Next, we observe that (s, a) succeeds, moving the agent to the intended goal state, $s_g(a)$. The distribution, $T(s, a)$, becomes non-trivial, a non-zero KL divergence is computed, and thus $R(s, a)$ gets an optimistically initialized reward, which does not depend on the size of the state-action space. **Algorithm 2** describes the steps of this optimistic initialization, and **Table 1** shows how $T(s, a)$ and $R(s, a)$ develop throughout the initialization process.

The distributions T , as initialized above, are compact and parsimonious, and they faithfully represent the most likely outcomes of the actions. Moreover, the second initialization step yields a non-zero KL Divergence, which is not sensitive to the

size of the state space. Importantly, the fact that our initialization of the state transition probabilities provides an initial measure of *interestingness* for each state-action allows us, *without choosing parameters*, to optimistically initialize the reward matrix with well defined *intrinsic rewards*. Consequently, we can employ a greedy policy, and aggressively explore the state-action space while focusing extra attention on the most *interesting* regions. As the curious agent explores, the intrinsic rewards decay in a logical way. A state-action, which deterministically leads to its goal state (**Table 2**) is less interesting over time than a state-action that leads to some other state (**Table 3**), and of course most *interesting* are state-actions with more possible outcomes (**Table 4**).

Algorithm 2: Curious_Explore($S, A, T, R, \gamma, \delta$)

```

begin
  for each state-action  $(s \in S, a \in A(s))$  do
    Observe( $s, a, s, T(s, a), R(s, a)$ )
    Observe( $s, a, s_g(a), T(s, a), R(s, a)$ )
  end
  while true do
    Value_Iteration( $S, A, T, R, \gamma, \delta$ )
     $s \leftarrow s_j | q(t_{before}) \in s_j$ 
     $a_{greedy} \leftarrow a | V(s, a) = argmax(\{V(s, a) | a \in A(s)\})$ 
    run  $a_{greedy}$  on the robot
     $s' \leftarrow s_j | q(t_{after}) \in s_j$ 
    Observe( $s, a, s', T(s, a), R(s, a)$ )
  end
end

```

Algorithm 3: Value_Iteration($S, A, T, R, \gamma, \delta$)

```

begin
  for each state-action  $(s \in S, a \in A(s))$  do
     $| V(s, a) \leftarrow 0.0$ 
  end
  for each state  $s \in S$  do
     $| V(s) \leftarrow 0.0$ 
  end
  while true do
    max_delta  $\leftarrow 0.0$ 
    for each state-action  $(s \in S, a \in A(s))$  do
       $| V_{new}(s, a) \leftarrow R(s, a) + \gamma \sum_{s'} T(s, a, s') V(s')$ 
      if  $V_{new}(s, a) - V(s, a) > max\_delta$  then
         $| max\_delta \leftarrow V_{new}(s, a) - V(s, a)$ 
      end
       $| V(s, a) \leftarrow V_{new}(s, a)$ 
    end
    for each state  $s \in S$  do
       $| V(s) \leftarrow argmax(\{V(s, a) | i = s\})$ 
    end
    if  $max\_delta < \delta$  then
      break
    end
  end
end

```

2.4. REINFORCEMENT LEARNING

At the beginning of section 2, we made the claim that a PRM planner's compact, incrementally expandable representation of known motions makes it a likely antecedent to a developmental learning system. Furthermore, we observed that many of the

Table 1 | Initialization of state transition probabilities.

Observation	T	P	$R = D_{KL}(P T)$
-	{}	{}	-
s_i	{1}	{2}	0
$s_g(a)$	{2,1}	{2,2}	0.0589

Table 2 | A predictable action ends in the predicted state.

Observation	T	P	$R = D_{KL}(P T)$
init	{2,1}	{2,2}	0.0589
$s_g(a)$	{2,2}	{2,3}	0.0201
$s_g(a)$	{2,3}	{2,4}	0.0095
$s_g(a)$	{2,4}	{2,5}	0.0052

Table 3 | A predictable action ends in a surprising state.

Observation	T	P	$R = D_{KL}(P T)$
init	{2,1}	{2,2}	0.0589
s_j	{2,2,1}	{2,2,2}	0.0487
s_j	{2,2,2}	{2,2,3}	0.0196
s_j	{2,2,3}	{2,2,4}	0.0103

Table 4 | An unpredictable action.

Observation	T	P	$R = D_{KL}(P T)$
init	{2,1}	{2,2}	0.0589
s_a	{2,2,1}	{2,2,2}	0.0487
s_b	{2,2,2,1}	{2,2,2,2}	0.0345
s_c	{2,2,2,2,1}	{2,2,2,2,2}	0.0283
$s_g(a)$	{2,2,2,2,2}	{2,3,2,2,2}	0.0142
s_a	{2,3,2,2,2}	{2,3,3,2,2}	0.0133
s_b	{2,3,3,2,2}	{2,3,3,3,2}	0.0125

weaknesses of PRMs can be avoided by *embodimenting the planner* and coupling it to a low-level reactive controller. Proxied by this low-level controller, the planner is empowered to try out arbitrary control signals, however, it does not necessarily know what will happen. Therefore, the PRM's original model of the robot's state-action space, a simple graph, is insufficient, and a more powerful, probabilistic model, an MDP is required. Thus, modeling the robot-workspace system using an MDP arises naturally from the effort to improve the robustness of a PRM planner, and accordingly, Model-Based RL is the most appropriate class of learning algorithms to operate on the MDP.

Having specified what *action* means in terms of robot control (section 2.1), described the layout and meaning of the state-action space (section 2.2), and defined the way in which intrinsic reward is computed according to the AC principal (section 2.3), we are ready to incorporate these pieces in a Model-Based RL system, which develop into a path planner as follows: Initially, sets of states and actions will be chosen, according to some heuristic(s), such that the robot's configuration space is reasonably well covered and the RL computations are tractable. Then, the state transition probabilities will be learned for each state-action pair, as the agent explores the MDP by moving the robot about. This exploration for the purposes of model learning will be guided entirely by the intrinsic reward defined in section 2.3, and the curious agent will continually improve its model of the iCub and its configuration space. In order to exploit the planner, an external reward must be introduced, which can either be added to or replace the intrinsic reward function.

The MDP, which constitutes the path planner, is a tuple, $\langle S, A, T, R, \gamma \rangle$, where S is a finite set of m states, A is a finite set of actions, T is a set of state transition probability distributions, R is a reward function, and γ is a discount factor, which represents the importance of future rewards. This MDP is somewhat unusual in that not all of the actions $a \in A$ are available in every state $s \in S$. Therefore, we define sets, $A(s)$, which comprise the actions $a \in A$ that are available to the agent when it finds itself in state s , and $A = \bigcup_{s=1}^m A(s)$. The set of state transition probabilities becomes $T : \bigcup_{s=1}^m A(s) \times S \rightarrow [0, 1]$, and in general, the reward function becomes $R : \bigcup_{s=1}^m A(s) \times S \rightarrow \mathbb{R}$, although the intrinsic reward, $R_{\text{intrinsic}} : \bigcup_{s=1}^m A(s) \rightarrow \mathbb{R}$, varies only with state-action pairs (s, a) , as opposed to state-action-state triples (s, a, s') . The state transition probabilities, T , are learned by curious exploration (**Algorithm 2**, $\gamma = 0.9$, $\delta = 0.001$), the RL algorithm employed is value iteration (**Algorithm 3**), and the intrinsic reward is computed as shown in **Algorithm 1**.

3. RESULTS

Here we present the results of two online learning experiments. The first one learns a motion planner for a single limb, the iCub's arm, operating in an unobstructed workspace, while other body parts remain motionless. The planner must contend with self-collisions, and infeasibilities due to the relative lengths of the cables, which move the shoulder joints. These constraints are

static, in that they represent properties of the robot itself, which do not change regardless of the configuration of the workspace. Due to the static environment, a PRM planner would in principle be applicable, and the experiment provides a context in which to compare and contrast the PRM versus MDP planners. Still, the primary question addressed by this first experiment is: "To what extent does AC help the agent learn the state transition probabilities for the MDP planner in this real-world setting?"

In the second experiment, the iCub is positioned at a work table, which constitutes a large obstacle in its workspace. Three curious agents, unaware of one another's states, learn planners for the iCub's torso and two arms, respectively. One could in principle define a single curious MDP planner for the whole body, but this would result in an explosion of the state-action space such that running actual experiments on the iCub hardware would be prohibitively time consuming. The modular, parallel, multi-agent configuration of this second experiment is designed to address the question: "Can curious MDP planners scale to intelligently control the entire iCub robot?" And in observing the behavior emergent from the interactions between the 3 learners, this will be the question of primary importance. Also noteworthy, however, is that from the perspective of the arms, which do not know that the torso is moving, the table seems to be non-static. By analyzing the arm learning while disregarding the torso, one can gain insight into how the curious MDP planner copes with non-static environments, which would render the PRM planner inoperable.

3.1. PLANNING IN A STATIC ENVIRONMENT—LEARNING TO AVOID SELF-COLLISIONS AND CABLE LENGTH INFEASIBILITIES

In the first experiment, "Planning in a static environment," we compare the exploration of our artificially curious agent (AC), to two other agents using benchmark exploration strategies from the RL literature. One explores randomly (RAND), and the other always selects the state-action least tried (LT)⁵.

The state space is defined by choosing samples, which vary in 4 dimensions corresponding to three shoulder joints and the elbow. Each of these joints is sampled at 25%, 50%, and 75% of its range of motion, resulting in a 4D hyper-lattice with 81 vertices, which are connected to their $2^4 = 16$ nearest neighbors as per section 2.2.1, yielding $81 \times 16 = 1296$ state-actions. The intuition behind this choice of state space it comprises a compact yet reasonably well dispersed set of pre-reach poses.

The task is to find the infeasible region(s) of the configuration space, and learn the according state transition probabilities such that the agent can plan motions effectively. The task is relatively simple, but it is none the less a crucial aspect of any path planning that should take place on the iCub. Without deliberately avoiding self-collisions and cable length infeasibilities, a controller can and will break the iCub's cables, rendering it inoperable.

In comparing the AC agent with the RAND agent and the LT agent, we find that AC produces, by far, the best explorer

⁵If there are multiple least-tried state-actions (for example when none have been tried), a random one from the least tried set is selected.

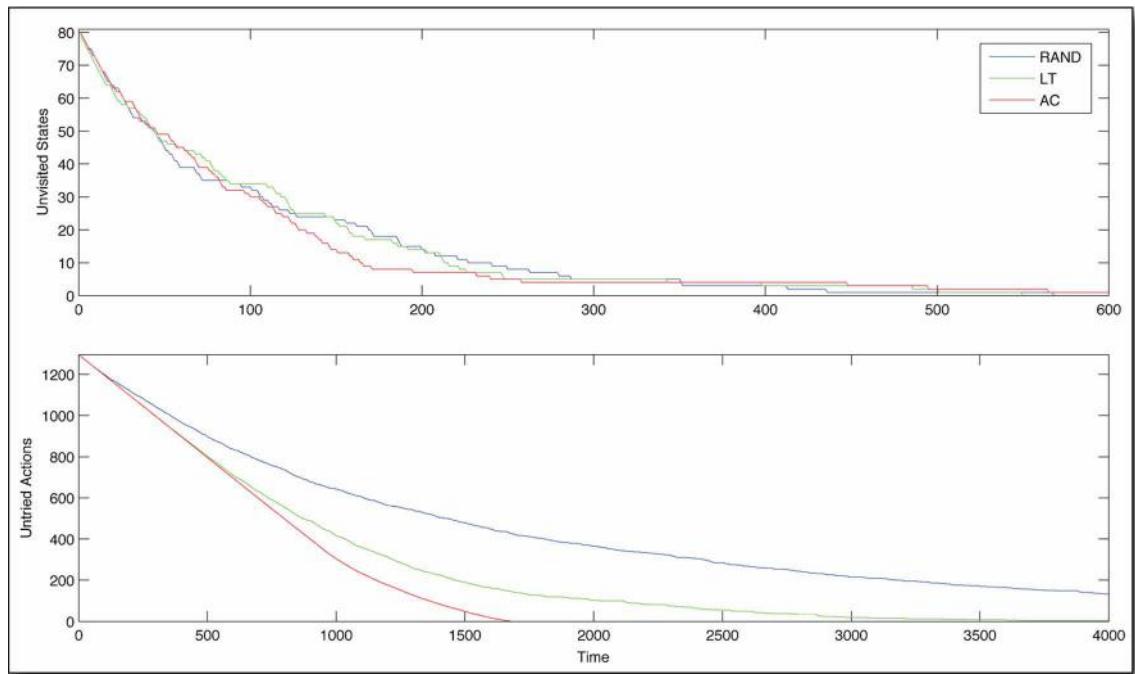


FIGURE 4 | State-action space coverage during early learning. The policy based on Artificial Curiosity (AC) explores the state-action space most efficiently, compared to policies based on random exploration (RAND) and always selecting the least tried state-action (LT). Time is measured in state transitions.

(Figure 4). In the early stages of learning, AC and LT try *only* novel actions, whereas RAND tries some actions repeatedly. Early on (before the agent has experienced about 220 state transitions), the only difference evident between AC and LT is that AC visits novel states more aggressively. This is intuitive upon reflection, as AC values states with *many* untried state-actions, and will traverse the state space to go find them, whereas LT has no global knowledge and just chooses the locally least tried state-action, regardless of where it leads. As learning continues, this key difference between AC and LT also begins to manifest in terms of the coverage of the action space. In fact, AC tries all possible state-actions in about $\frac{1}{2}$ the time it takes LT.

Moving on to the tabulated number of times that each state was visited and each state-action was tried, after 4000 state transitions, again we see that AC exhibits preferable behavior to LT and RAND (Figure 5). AC results in distributions of visits over states and tries over state-actions, which are more uniform than those resultant of RAND and LT. Moreover, we see a number of large spikes, where the AC agent became very interested in certain state-actions. In fact, these are the actions that run the robot into its constraints, and therefore do not cause the anticipated state transition (Equation 4). While most of the state-actions' rewards decay according to Table 2, these spikes were generated by state-actions whose rewards are governed by Table 3, and they are thus more *interesting* to the agent.

The decay of the intrinsic reward over the state-action space over time is shown in Figure 6. The uniformity of the decay is intuitive, since whenever there exists a spike in the reward function, the AC agent goes and gets it, thereby gaining experience and

decrementing the reward for the state-action tried. Thus, differing rates of decay (Tables 1–4) govern the frequency with which the agent tries the different state-actions.

The learned MDP is pictured in Figure 7. Since the workspace of the arm is unobstructed, most of the state-actions behave as expected, reliably taking the agent to the intended goal state (Equation 3). These deterministic state-actions, shown in gray, are *boring*. The *interesting* ones, each shown in a different color, took the agent to a novel state, which was not represented in the initial state transition distribution for that state-action. Since the environment is static, one would expect even these novel state transitions to be deterministic (Equation 4), and some of them are (red, yellow, purple, light blue). However, the other state-actions (green, brown, and dark blue) sometimes lead to the intended goal state and sometimes lead to one other state, despite the static constraints and the fact that each state-action always runs the same control code.

The fact that static constraints do not necessarily lead to deterministic state transitions is quite interesting. It shows that the iCub, an advanced, light-weight, cable-driven robot, exhibits important non-linearities, due to its mechanics and/or embedded control systems, which prevent it from reliably and repeatably executing precise motions. Therefore, a *plan first, act later* approach, such as PRM planning, will never work well on robots such as the iCub. Plans will sometimes fail at runtime, and not necessarily in a repeatable manner. In fact the lighter and more flexible robots get, the more non-linearities will dominate their dynamics, which is an important motivation for continuing to

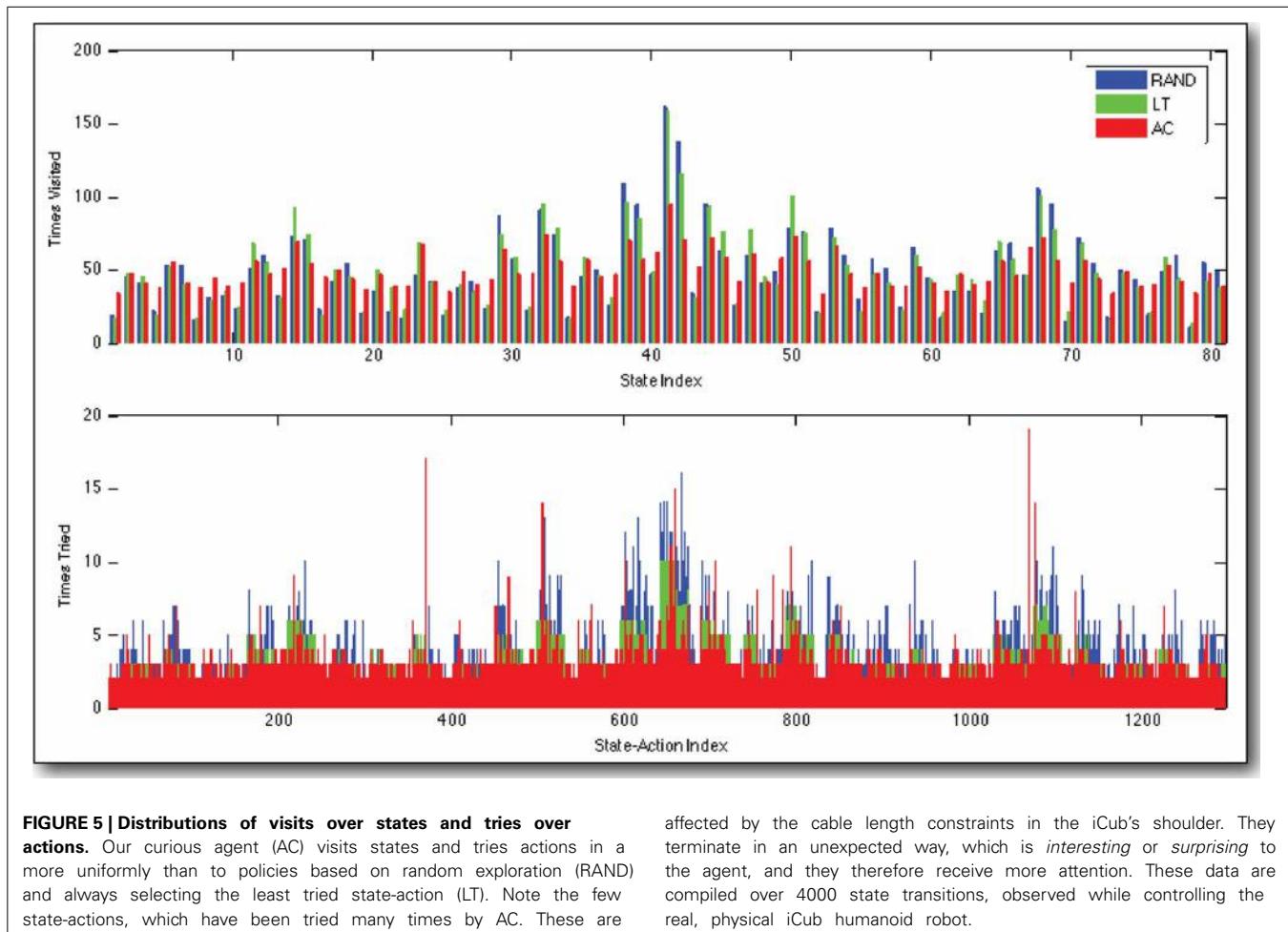


FIGURE 5 | Distributions of visits over states and tries over actions. Our curious agent (AC) visits states and tries actions in a more uniformly than to policies based on random exploration (RAND) and always selecting the least tried state-action (LT). Note the few state-actions, which have been tried many times by AC. These are

affected by the cable length constraints in the iCub's shoulder. They terminate in an unexpected way, which is *interesting* or *surprising* to the agent, and they therefore receive more attention. These data are compiled over 4000 state transitions, observed while controlling the real, physical iCub humanoid robot.

develop more robust solutions, such as the MDP motion planning presented here.

3.2. DISCOVERING THE TABLE WITH A MULTI-AGENT RL SYSTEM

In the second experiment, we control both of the iCub's arms and its torso, 12 DOF in total. A hypercube in 12 dimensions has 4096 vertices, and a rank 3 hyper-lattice has 531,441 vertices. Clearly, uniform sampling in 12 dimensions will not yield a feasible RL problem. Therefore, we have parallelized the problem, employing three curious agents that control each arm and the torso separately, not having access to one another's state. The state-action spaces for the arms are exactly as described in the previous experiment, and the state-action space for the 3D torso is defined in an analogous manner (25%, 50%, and 75% of each joint's range of motion), resulting in a 3D lattice with 27 vertices, which are connected to their $2^3 = 8$ nearest neighbors as per section 2.2.1, yielding $27 \times 8 = 216$ state-actions.

We place the iCub in front of a work table, and all three learners begin exploring (Figure 8). The three agents operate strictly in parallel, having no access to any state information from the others, however, they are loosely coupled through their effects on the robot. For example, the operational space position of the hand (and therefore whether or not it is colliding with the table)

depends not only on the positions of the joints in the arm, but also on the positions of the joints in the torso. Thus, we have three interacting POMDPs, each of which has access to a different piece of the complete robot state, and the most *interesting* parts of the state-action spaces are where the state of one POMDP affects some state transition(s) of another.

When the torso is upright, each arm can reach all of the states in its state space, but when the iCub is bent over at the waist, the shoulders are much closer to the table, and some of the arms' state-actions become infeasible, because the robot's hands hit the table. Such interactions between the learners produce state-transition distributions, like the one shown in Figure 9, which are much richer than those from the previous experiment. Moreover these state-actions are the most interesting because they generate the most slowly decaying intrinsic reward of the type shown in Table 4. The result is that the arms learn to avoid constraints as in the first experiment, but over time, another behavior emerges. The iCub becomes interested in the table, and begins to touch it frequently. Throughout the learning process, it spends periods of time exploring, investigating its static arm constraints, and touching the table, in a cyclic manner, as all the intrinsic rewards decay over time in a manner similar to Figure 6.

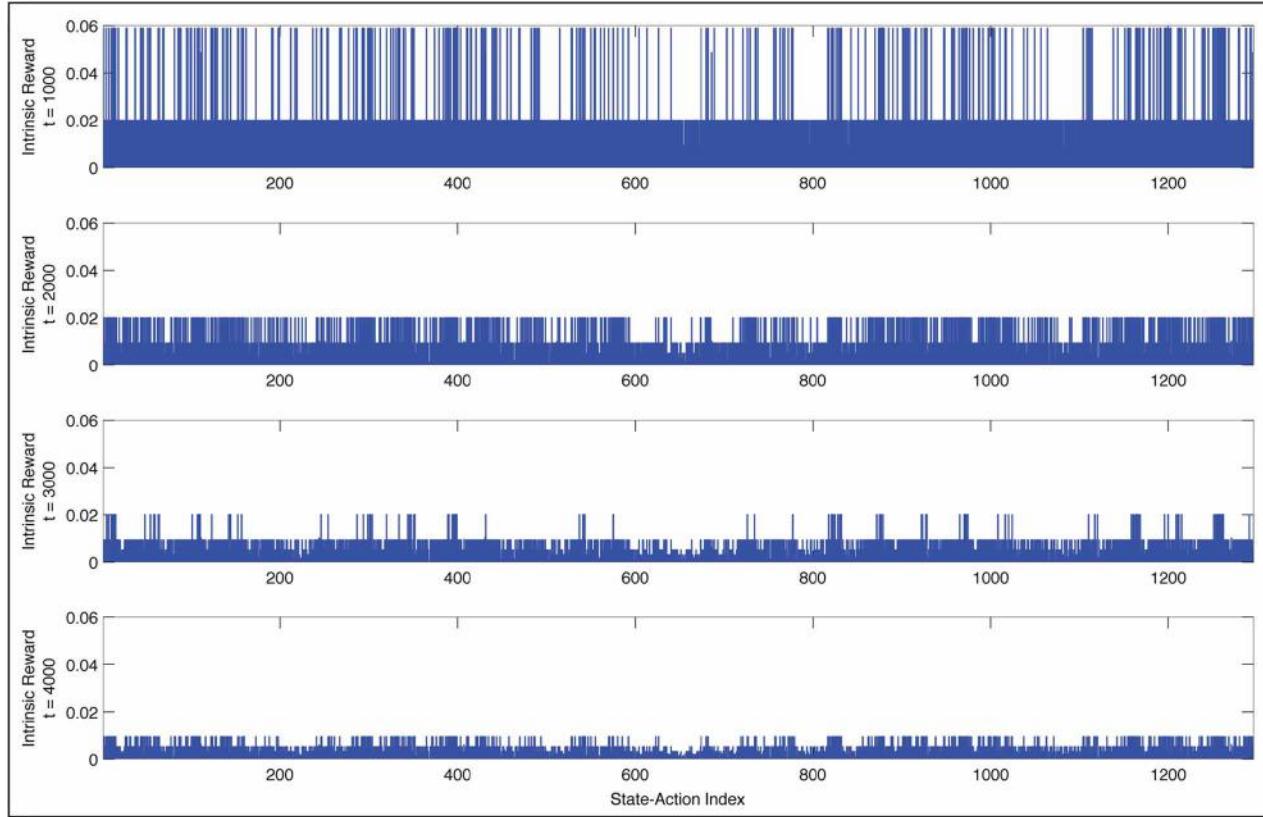


FIGURE 6 | Decay of intrinsic reward over time. These snapshots of the reward distribution state-actions (x-axis) over time (from **top** to **bottom**) show how our curious agent becomes *bored* as it builds a better and better model of the state-action space. Time is measured in state transitions.

In **Figure 10**, we have tabulated the distribution of tries over the state-action space for each of the three learners after 18,000 state transitions, or a little more than two full days of learning. As in the previous experiment, we see that the curious agent prefers certain state-actions, selecting them often. Observing the behavior of the robot during the learning process, it is clear that these frequently chosen state-actions correspond to putting the arm down low, and leaning forward, which result in the iCub's hand interacting with the table. Furthermore, the distribution of selected state-actions for the right arm and the left arm are very similar indeed. This is to be expected, since the arms are mechanically very similar and their configuration spaces have been discretized the same way. It is an encouraging result, which seems to indicate that the variation in the number of times different state-actions are selected does indeed capture the extent to which those state-actions interfere with (or *are* interfered with by) the other learners.

The emergence of the table exploration behavior is quite promising with respect to the ultimate goal of using MDP based motion planning to control an entire humanoid *intelligently*. We partitioned an intractable configuration space into several loosely coupled RL problems, and with only intrinsic rewards to guide their exploration, the learning modules coordinated their behavior, causing the iCub to explore the surface of the work table

in front of it. Although the state spaces were generated using a coarse uniform sampling, and the object being explored was large and quite simple, the experiment nevertheless demonstrates that MDP motion planning with AC can empower a humanoid robot with many DOF to explore its environment in a structured way and build useful, reusable models.

3.2.1. Planning in a dynamic environment

There is an alternative way to view the multi-agent experiment. Because the arm does not have access to the torso's state, the experiment is exactly analogous to one in which the arm is the only learner and the table is a dynamic obstacle, moving about as the arm learns. Even from this alternative viewpoint, it is none the less true that some actions will have different outcomes, depending on the table configuration, and will result in state transition distributions like the one shown in **Figure 9**. The key thing to observe here is that if we were to exploit the planner by placing an external reward at some goal, removing the intrinsic rewards, and recomputing the value function, then the resulting policy/plan will try to avoid the unpredictable regions of the state-action space, where state transition probabilities are relatively low. In other words, training an MDP planner in an environment with dynamic obstacles, will produce policies that plan around regions where there tend to be obstacles.

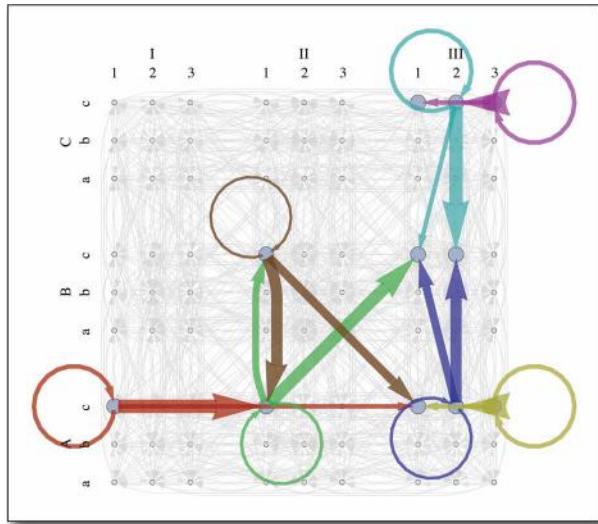


FIGURE 7 | The learned single-arm MDP planner. The 4D state space is labeled as follows: shoulder flexion/extension (1,2,3), arm abduction/adduction (a,b,c), lateral/medial arm rotation (I,II,III), elbow flexion/extension (A,B,C). Each color represents an *interesting* state-action, which often takes the agent to some unexpected state. Each arrow of a particular color represents a state transition probability and the weight of the arrow is proportional to the magnitude of that probability. Arrows in gray represent *boring* state-actions. These work as expected, reliably taking the agent to the intended goal state, to which they point.

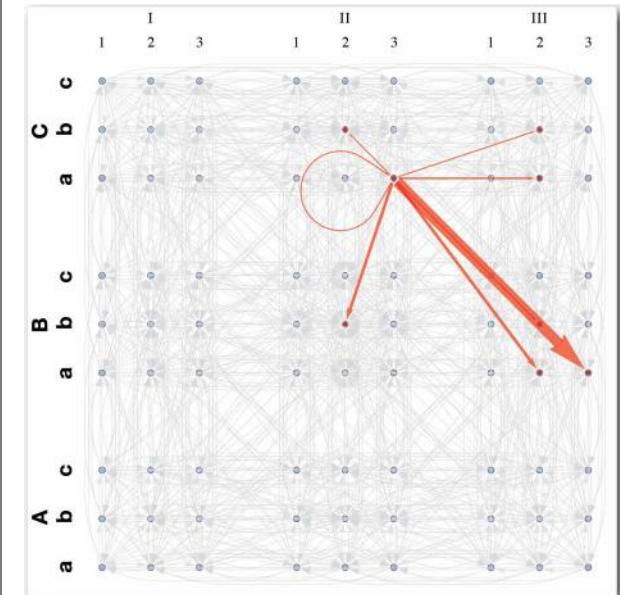


FIGURE 9 | State space and transition distribution for an interesting arm action in multi-agent system. The 4D state space is labeled as follows: shoulder flexion/extension (1,2,3), arm abduction/adduction (a,b,c), lateral/medial arm rotation (I,II,III), elbow flexion/extension (A,B,C). The red arrows show the distribution of next states resultant of an *interesting* state-action, which causes the hand to interact with the table. Each arrow represents a state transition probability and the weight of the arrow is proportional to the magnitude of that probability. Arrows in gray represent *boring* state-actions. These work as expected, reliably taking the agent to the intended goal state, to which they point.



FIGURE 8 | Autonomous exploration. This composite consists of images taken every 30 s or so over the first hour of the experiment described in section 3.2.1. Although learning has just begun, we already begin to see that the cloud of robot poses is densest (most opaque) near the table. Note that the compositing technique as well as the wide angle lens used here create the illusion that the hands and arms are farther from the table than they really are. In fact, the low arm poses put the hand or the elbow within 2 cm of the table, as shown in Figure 8.

4. DISCUSSION

In this paper we have developed an embodied, curious agent, and presented the first experiments we are aware of, which exploit AC to learn an MDP based motion planner for a real, physical

humanoid robot. We demonstrated the efficacy of the AC concept with a simple learning experiment wherein one learner controls one of the iCub humanoid's arms. The primary result of this first experiment was that the iCub's autonomous exploratory behavior, guided by AC, efficiently generated a continually improving Markov model, which can be (re)used at any time to quickly satisfy path planning queries.

Furthermore, we conducted a second experiment, in which the iCub was situated at a work table while three curious agents controlled its arms and torso, respectively. Acting in parallel, the three agents had no access to one another's state, however, the interaction between the three learners produced an interesting emergent behavior; guided only by intrinsic rewards, the torso and arm coordinated their movements such that the iCub explored the surface of the table.

4.1. SCALABILITY

From the standpoint of scalability, the state spaces used for the arms and torso were more than tractable. In fact the time it took the robot to move from one pose/state to another exceeded the time it took to update the value function by approximately an order of magnitude. From an experimental standpoint, the limiting factor with respect to the size of the state-action space was the time it took to try all the state-actions a few times. In these experiments we connected states to their 2^n nearest neighbors, where n

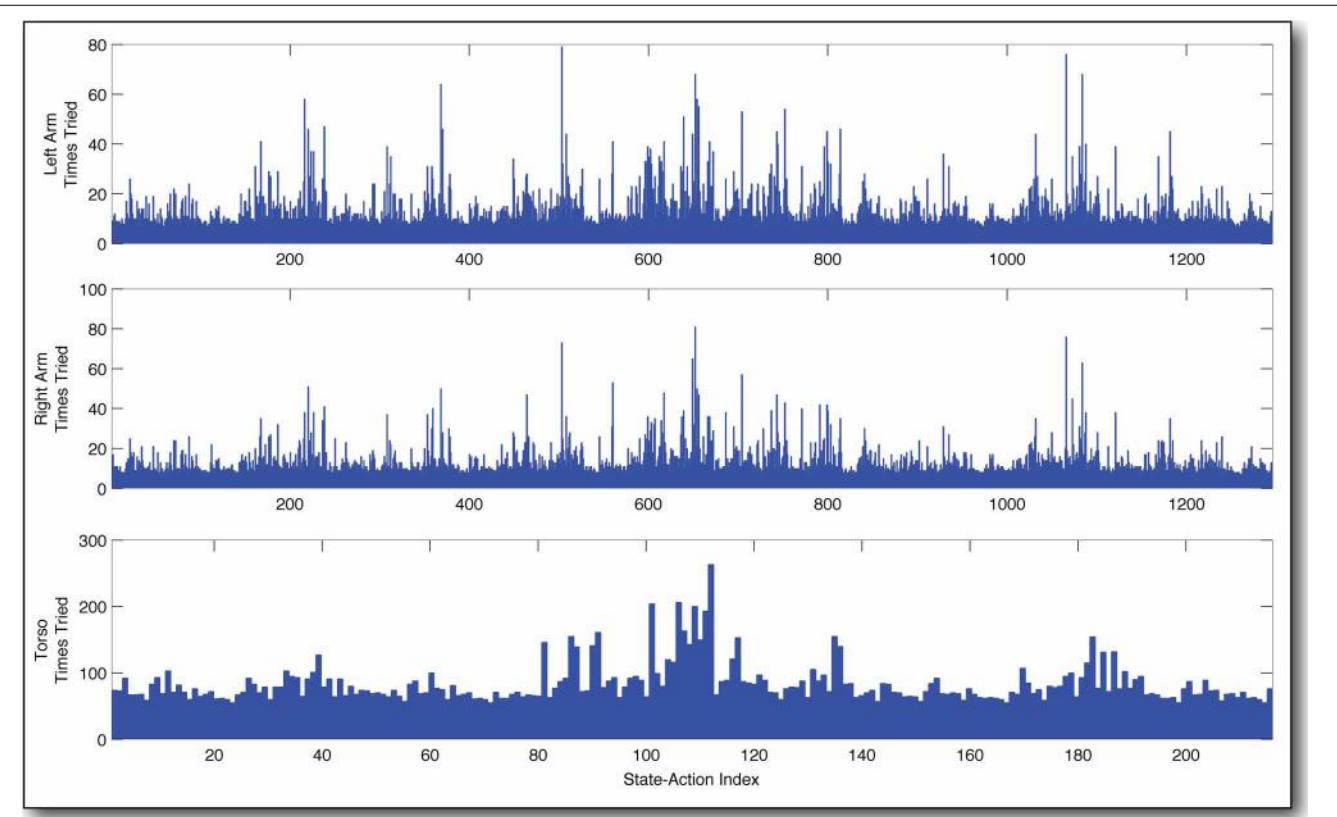


FIGURE 10 | Frequency of actions taken by three curious agents in parallel. The most *interesting* actions are selected much more often than the others. They correspond to moving the arm down and

leaning the torso forward. This results in the iCub robot being interested in the table surface. Note the similarity in the behavior of the two arms.

is the dimensionality of the configuration space, and we ran the learning experiments for some 12 and 50 h, respectively.

Increasing the number of states in the MDPs would undoubtedly yield a more powerful planner, but it would also increase the time required to learn the models sufficiently. One way to mitigate this effect would be to reduce the number of connections between states. In fact, our impression from qualitative observation of the learning process is that the connectivity of the state space was denser than necessary. Alternatively, we could of course allow the robot to learn for longer. After all, children require years to learn the kinds of skills we are trying to replicate.

4.2. DIVERSITY OF ACTIONS

In these experiments, the implementation of actions (section 2.1) was designed to facilitate motion planning for the purpose of avoiding non-linear constraints on the robot configuration such as such as unwanted collisions. The actions simply set an attractor in configuration space via the MoBeE framework at the Voronoi center of a region of configuration space, which defines a state. The robot then moved according to the transient response of the dynamical system within MoBeE. The result was that our MDP functioned as a sort of enhanced version of a PRM planner, however, the RL framework presented here is in principal capable of much more.

In addition to the position control presented here, our MoBeE framework supports force control in both joint space and operational space, and as far as our RL implementation is concerned, actions can contain arbitrary control code. Therefore, future curious agents for the iCub will benefit from different action modalities, such as operational space reaches or even learned dynamic motion primitives (Schaal et al., 2005).

4.3. BOOTSTRAPPING THE STATE SPACE

In our view, the main shortcoming of the work presented here is that we have constructed the state-action spaces by hand. In the future, it would be greatly desirable to automate this process, perhaps in the form of an offline process that can run in the background, searching for sets of interesting poses (Stollenga et al., 2013), and incrementally expanding the state-action space. The only part of this proposition, which is unclear, is how to evaluate the quality of the samples that should potentially define new states.

4.4. HIERARCHIES OF AGENTS

The experiment “Discovering the table” is promising with respect to the goal of extending our multi-agent MDP motion planning to hierarchies of agents. The *interesting* (most frequently selected) state-actions, as discovered by the current system, constitute each

agent's ability to interact with the others. Therefore they are exactly the actions that should be considered by a parent agent, whose job it would be to coordinate the different body parts. It is our strong suspicion that all state-actions, which are not interesting to the current system, can be compressed as "irrelevant" in the eyes of such a hypothetical parent agent. However, to develop the particulars of the communication up and down the hierarchy remains a difficult challenge, and the topic of ongoing work.

ACKNOWLEDGMENTS

The authors would like to thank Jan Koutnik, Matt Luciw, Tobias Glasmachers, Simon Harding, Gregor Kaufmann, and Leo Pape, for their collaboration, and their contributions to this project.

FUNDING

This research was supported by the EU Project IM-CLeVeR, contract no. FP7-IST-IP-231722.

REFERENCES

- Barto, A. G., Singh, S., and Chentanez, N. (2004). "Intrinsically motivated learning of hierarchical collections of skills," in *Proceedings of the 3rd International Conference Development Learning* (San Diego, CA), 112–119.
- Barto, A. G., Sutton, R. S., and Anderson, C. W. (1983). Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE Trans. Syst. Man Cybern. SMC-13*, 834–846. doi:10.1109/TSMC.1983.6313077
- Brooks, R. (1991). Intelligence without representation. *Artif. Intell.* 47, 139–159. doi:10.1016/0004-3702(91)90053-M
- D'Souza, A., Vijayakumar, S., and Schaal, S. (2001). Learning inverse kinematics. *Int. Conf. Intell. Robots Syst.* 1, 298–303. doi:10.1109/IROS.2001.973374
- Fedorov, V. V. (1972). *Theory of Optimal Experiments*. New York, NY: Academic Press.
- Frank, M., Leitner, J., Stollenga, M., Kaufmann, G., Harding, S., Forster, A., et al. (2012). "The modular behavioral environment for humanoids and other robots (mobeey)," in *9th International Conference on Informatics in Control, Automation and Robotics (ICINCO)*.
- Gordon, G., and Ahissar, E. (2011). "Reinforcement active learning hierarchical loops," in *The 2011 International Joint Conference on Neural Networks (IJCNN)* (San Jose, CA), 3008–3015. doi:10.1109/IJCNN.2011.6033617
- Gordon, G., and Ahissar, E. (2012). "A curious emergence of reaching," in *Advances in Autonomous Robotics, Joint Proceedings of the 13th Annual TAROS Conference and the 15th Annual FIRA RoboWorld Congress*, (Bristol: Springer Berlin Heidelberg), 1–12. doi:10.1007/978-3-642-32527-4_1
- Huang, X., and Weng, J. (2002). *Novelty and reinforcement learning in the value system of developmental robots*. eds. C. G. Prince, Y. Demiris, Y. Marom, H. Kozima and C. Balkenius (Lund University Cognitive Studies), 47–55. Available online at: <http://cogprints.org/2511/>
- Iossifidis, I., and Schoner, G. (2004). "Autonomous reaching and obstacle avoidance with the anthropomorphic arm of a robotic assistant using the attractor dynamics approach," in *Proceedings ICRA'04 2004 IEEE International Conference on Robotics and Automation*. Vol. 5 (IEEE, Bochum, Germany), 4295–4300. doi:10.1109/ROBOT.2004.1302393
- Iossifidis, I., and Schoner, G. (2006). Reaching with a redundant anthropomorphic robot arm using attractor dynamics. *VDI BERICHTE* 1956, 45.
- Itti, L., and Baldi, P. F. (2005). "Bayesian surprise attracts human attention," in *Advances in neural information processing systems (NIPS)*, 547–554.
- Kaelbling, L. P., Littman, M. L., and Moore, A. W. (1996). Reinforcement learning: a survey. *J. Artif. Intell. Res.* 4, 237–285.
- Khatab, O. (1986). Real-time obstacle avoidance for manipulators and mobile robots. *Int. J. Rob. Res.* 5, 90.
- Kim, J., and Khosla, P. (1992). Real-time obstacle avoidance using harmonic potential functions. *IEEE Trans. Rob. Automat.* 8, 338–349.
- Kompella, V., Luciw, M., Stollenga, M., Pape, L., and Schmidhuber, J. (2012). "Autonomous learning of abstractions using curiosity-driven modular incremental slow feature analysis," in *Proceedings of the Joint International Conference Development and Learning and Epigenetic Robotics (ICDL-EPIROB-2012)* (San Diego, CA). doi:10.1109/DevLrn.2012.6400829
- Latombe, J., Kavraki, L., Svestka, P., and Overmars, M. (1996). Probabilistic roadmaps for path planning in high-dimensional configuration spaces. *IEEE Trans. Rob. Automat.* 12, 566–580. doi: 10.1109/70.508439
- LaValle, S. (1998). Rapidly-exploring random trees: a new tool for path planning. Technical report, Computer Science Department, Iowa State University.
- LaValle, S. (2006). *Planning Algorithms*. Cambridge, MA: Cambridge University Press. doi: 10.1017/CBO9780511546877
- Li, T., and Shie, Y. (2007). An incremental learning approach to motion planning with roadmap management. *J. Inf. Sci. Eng.* 23, 525–538.
- Lindley, D. V. (1956). On a measure of the information provided by an experiment. *Annal. Math. Stat.* 27, 986–1005. doi: 10.1214/aoms/1177728069
- Luciw, M., Graziano, V., Ring, M., and Schmidhuber, J. (2011). "Artificial curiosity with planning for autonomous perceptual and cognitive development," in *IEEE International Conference on Development and Learning (ICDL)*, vol. 2 (Frankfurt am Main), 1–8. doi: 10.1109/DEVLRN.2011.6037356
- Lungarella, M., Metta, G., Pfeifer, R., and Sandini, G. (2003). Developmental robotics: a survey. *Connect. Sci.* 15, 151–190. doi: 10.1080/09540090310001655110
- Metta, G., Sandini, G., Vernon, D., Natale, L., and Nori, F. (2008). "The icub humanoid robot: an open platform for research in embodied cognition," in *Proceedings of the 8th Workshop on Performance Metrics for Intelligent Systems* (New York, NY: ACM), 50–56.
- Mugan, J., and Kuipers, B. (2012). Autonomous learning of high-level states and actions in continuous environments. *IEEE Trans. Auton. Mental Dev.* 4, 70–86. doi: 10.1109/TAMD.2011.2160943
- Ngo, H., Luciw, M., Foerster, A., and Schmidhuber, J. (2012). "Learning skills from play: artificial curiosity on a katana robot arm," in *Proceedings of the 2012 International Joint Conference of Neural Networks (IJCNN)* (Brisbane, Australia).
- Nori, F., Natale, L., Sandini, G., and Metta, G. (2007). "Autonomous learning of 3d reaching in a humanoid robot," in *IEEE/RSJ International Conference on Intelligent Robots and Systems* (San Diego, CA), 1142–1147.
- Oudeyer, P.-Y., Kaplan, F., and Hafner, V. V. (2007). Intrinsic motivation systems for autonomous mental development. *IEEE Trans. Evol. Comput.* 11, 265–286. doi: 10.1109/TEVC.2006.890271
- Pape, L., Oddo, C. M., Controzzi, M., Cipriani, C., Förster, A., Carrozza, M. C., et al. (2012). Learning tactile skills through curious exploration. *Front. Neurorobot.* 6:6. doi: 10.3389/fnbot.2012.00006
- Perez, A., Karaman, S., Shkolnik, A., Frazzoli, E., Teller, S., and Walter, M. (2011). "Asymptotically-optimal path planning for manipulation using incremental sampling-based algorithms," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (San Francisco, CA: IEEE), 4307–4313. doi: 10.1109/IROS.2011.6094994
- Peters, J., and Schaal, S. (2008). Learning to control in operational space. *Int. J. Rob. Res.* 27, 197. doi: 10.1177/0278364907087548
- Piaget, J., and Cook, M. T. (1952). *The Origins of Intelligence in Children*. New York, NY: International Universities Press. doi: 10.1037/11494-000
- Schaal, S., Peters, J., Nakanishi, J., and Ijspeert, A. (2005). "Learning movement primitives," in *International Symposium on Robotics Research*, Vol. 15, eds. D. Paolo and C. Raja (Berlin Heidelberg: Springer), 561–572. doi: 10.1007/11008941_60
- Schmidhuber, J. (1991a). "Curious model-building control systems," in *Proceedings of the International Joint Conference on Neural Networks*. Vol. 2 (Singapore: IEEE Press), 1458–1463.
- Schmidhuber, J. (1991b). "A possibility for implementing curiosity and boredom in model-building neural controllers," in *Proceedings of the International Conference on Simulation of Adaptive Behavior: From Animals to Animats*, eds J. A. Meyer and S. W. Wilson (Cambridge, MA: MIT Press/Bradford Books), 222–227.
- Schmidhuber, J. (2006). Developmental robotics, optimal artificial curiosity, creativity, music, and the fine arts. *Connect. Sci.* 18, 173–187. doi: 10.1080/09540090600768658
- Schmidhuber, J. (2013). POWERPLAY: training an increasingly general problem solver by continually searching for the simplest still unsolvable problem. *Front. Psychol.* 4:313. doi:10.3389/fpsyg.2013.00313
- Schoner, G., and Dose, M. (1992). A dynamical systems approach to task-level system integration used to plan and control autonomous vehicle motion. *Rob. Auton. Syst.* 10, 253–267. doi: 10.1016/0921-8890(92)90004-I

- Srivastava, R. K., Steunebrink, B. R., and Schmidhuber, J. (2013). First experiments with POWERPLAY. *Neural Netw.* 41, 130–136. doi: 10.1016/j.neunet.2013.01.022
- Stollenga, M., Pape, L., Frank, M., Leitner, J., Förster, A., and Schmidhuber, J. (2013). “Task-relevant roadmaps: a framework for humanoid motion planning,” in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. Tokyo.
- Storck, J., Hochreiter, S., and Schmidhuber, J. (1995). “Reinforcement driven information acquisition in non-deterministic environments,” in *Proceedings of the International Conference on Artificial Neural Networks*. Vol. 2 (Paris: Citeseer), 159–164.
- Sun, Z., Hsu, D., Jiang, T., Kurniawati, H., and Reif, J. H. (2005). Narrow passage sampling for probabilistic roadmap planning. *IEEE Trans. Rob.* 21, 1105–1115. doi: 10.1109/TRO.2005.853485
- Sutton, R., Precup, D., and Singh, S. (1999). Between MDPs and semi-MDPs: a framework for temporal abstraction in reinforcement learning. *Artif. Intell.* 112, 181–211. doi: 10.1016/S0004-3702(99)00052-1
- Sutton, R. S., and Barto, A. G. (1998). *Reinforcement Learning: An Introduction*, Vol. 1. Cambridge, MA: Cambridge Univ Press.
- Weng, J., McClelland, J., Pentland, A., Sporns, O., Stockman, I., Sur, M., et al. (2001). Autonomous mental development by robots and animals. *Science* 291, 599–600. doi: 10.1126/science.291.5504.599

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 05 July 2013; accepted: 04 December 2013; published online: 06 January 2014.

*Citation: Frank M, Leitner J, Stollenga M, Förster A and Schmidhuber J (2014) Curiosity driven reinforcement learning for motion planning on humanoids. *Front. Neurorobot.* 7:25. doi: 10.3389/fnbot.2013.00025*

*This article was submitted to the journal *Frontiers in Neurorobotics*.*

Copyright © 2014 Frank, Leitner, Stollenga, Förster and Schmidhuber. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



A psychology based approach for longitudinal development in cognitive robotics

J. Law*, P. Shaw, K. Earland, M. Sheldon and M. Lee

Department of Computer Science, Aberystwyth University, Aberystwyth, UK

Edited by:

Marco Mirolli, Istituto di Scienze e Tecnologie della Cognizione, Italy

Reviewed by:

Minoru Asada, Osaka University, Japan

Alex Pitti, University of Cergy Pontoise, France

***Correspondence:**

J. Law, Department of Computer Science, Aberystwyth University, Llandinam Building, Aberystwyth, Ceredigion SY23 3DB, UK
e-mail: jxl@aber.ac.uk

A major challenge in robotics is the ability to learn, from novel experiences, new behavior that is useful for achieving new goals and skills. Autonomous systems must be able to learn solely through the environment, thus ruling out *a priori* task knowledge, tuning, extensive training, or other forms of pre-programming. Learning must also be cumulative and incremental, as complex skills are built on top of primitive skills. Additionally, it must be driven by intrinsic motivation because formative experience is gained through autonomous activity, even in the absence of extrinsic goals or tasks. This paper presents an approach to these issues through robotic implementations inspired by the learning behavior of human infants. We describe an approach to developmental learning and present results from a demonstration of longitudinal development on an iCub humanoid robot. The results cover the rapid emergence of staged behavior, the role of constraints in development, the effect of bootstrapping between stages, and the use of a schema memory of experiential fragments in learning new skills. The context is a longitudinal experiment in which the robot advanced from uncontrolled motor babbling to skilled hand/eye integrated reaching and basic manipulation of objects. This approach offers promise for further fast and effective sensory-motor learning techniques for robotic learning.

Keywords: development, robotics, intrinsic motivation, staged learning, constraints

1. INTRODUCTION

The question of autonomy poses a particularly hard challenge for robotics research—how can robots grow through the “open-ended acquisition of novel behavior?” That is, given an embodied robot system with some primitive actions, how can it learn appropriate new behaviors to deal with new and novel experiences. It is apparent that this will involve the integration of past experience with new sensorimotor possibilities, but this remains a difficult, important, and unsolved research area. We report on experiments that illustrate the value of a developmental attack on this issue.

Developmental robotics is a recent field of study that recognizes the role of epigenetic development as a new paradigm for adaptation and learning in robotics. Most research in this field reports on specific topics in development such as motivation, embodiment, enactive growth, imitation, self-awareness, agent interaction and other issues. Such investigations are exploring effective modeling methods and increasing our understanding of the many and varied aspects of the phenomenon of development. For general principles and reviews see Lungarella et al. (2003); Asada et al. (2009); Stoytchev (2009).

In our research, presented here, we place emphasis on two key features: the role of psychological theories in development; and the importance of longitudinal studies.

While all work in this field takes account of current knowledge in both neuroscience and experimental psychology, there exists a significant lacuna between psychological theories of development and our ability to implement those theories as working developmental algorithms. There is a large body of experimental work in

psychology and we view psychological theory as a distillation of the understanding gained from such work that can guide modeling and help focus on key issues. In our work we are inspired by Piaget’s extensive studies, particularly his emphasis on: staged growth; the fundamental role of sensory-motor development; and his constructivist approach (Piaget, 1973).

While recognizing that longitudinal development is a central issue, much current research has been focused on topics at particular stages in development, often involving cross-sectional data. This means that correspondingly less attention is being paid to the cumulative effects of continuous growth and the totality of the developmental trajectory. Some significant studies on longitudinal aspects have resulted in various time-lines or roadmaps being produced. These translate the developmental progression seen in human infants into a suggested or plausible trajectory of behavioral competence that might be expected of a successful robot model. Examples of roadmaps include that from the iTalk project (Cangelosi et al., 2010), the broad approach of the Jst Erato Asada project (<http://jeap.jp/>), and the output from the RobotCub project (Vernon et al., 2010). The work reported here is based on a detailed infant timeline (Law et al., 2011).

This paper presents a model of longitudinal development analogous to part of the sensory-motor development of a human infant from birth to 6 months. **Figure 1** gives an overview of the development timeline we use which is fully described in Law et al. (2011). We use an iCub humanoid robot (Natale et al., 2013) as the platform for our experiment. The robot is given no prior abilities and the task is to learn to coordinate and gain control of

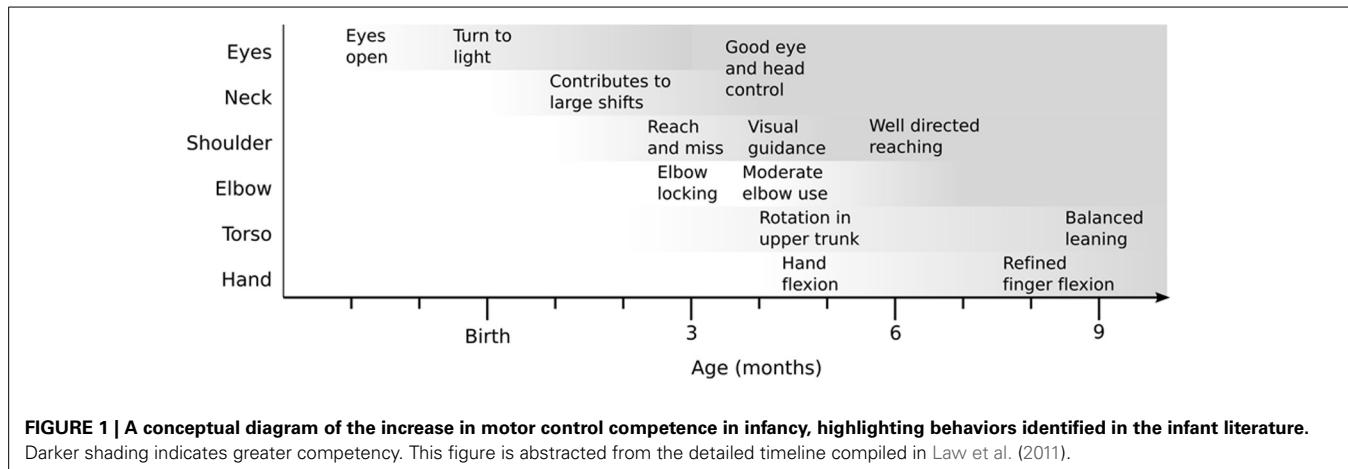


FIGURE 1 | A conceptual diagram of the increase in motor control competence in infancy, highlighting behaviors identified in the infant literature.

Darker shading indicates greater competency. This figure is abstracted from the detailed timeline compiled in Law et al. (2011).

the motor system through some form of exploratory process. The criteria for success is in achieving sufficient competence to visually detect objects, reach toward and grasp them, and move them around in the environment. In other words, the aim is to advance from no understanding of the structure of the sensory-motor hardware to achieving skilled hand-eye coordination, involving reaching skills and mastery of the local egocentric space. To enable this, we provide the robot with a suitable architecture, on which to learn sensorimotor coordination, and a series of constraints designed to shape learning along a trajectory similar to that seen in infancy.

In this paper we present results from a complete longitudinal experiment that shows a full developmental cycle, progressing through several distinct behavioral stages and increasing competence from essentially no control (random motor action) to skilled visio-integrated reaching and manipulation of objects. This experiment was made possible by guidance from the results of several investigations into the various subsystems involved: eyes, head, arms, etc. While there is insufficient space to expand on all these prior studies, we reference them where appropriate in order to provide further background on particular aspects of our architecture. Particular new contributions include the use and control of the torso, reaching for objects, and schema learning for novel actions. The key findings reported here include: evidence for the speed and effectiveness of staged behavior; evidence for the role of constraints in staged development; the effect of bootstrapping between stages; and the use of a schema memory of experiential fragments in learning new skills. These are seen in the context of a longitudinal sequence showing the development in a continuous process—to our knowledge, this has not been performed on an iCub robot previously.

2. MATERIALS AND METHODS

The experiments we describe here were performed on an iCub humanoid robot (Natale et al., 2013), depicted in terms of the sensor and motor systems of interest in **Figure 2**. The robot has a total of 53 independent degrees of freedom, however, here we only consider the 15 that are involved in hand/eye coordination (excluding the legs, hips, wrists and fingers). Although fine hand control (e.g., grasp adaptation to affordances) is not part of our

study we do use some of the wrist and finger motors for simple hand closing reflexes.

The robot has joint angle sensors to provide proprioception and touch sensors in the hand that can simulate a primitive tactile sense. The eyes are color CCD cameras and so provide two 2D images, but the center of the retinal image is taken to be the loci of interest and the two eyes converge on a fixation point in a 3D visual space¹. This visual space can be affected by several motor systems that cause bodily movement; for example if the head moves it will disturb the gaze point. However, the pattern of the disruption to vision is repeatable and lasting and so can be learned. This is shown in **Figure 2** as a mapping process resulting in the *gaze space*—the space of visual fixation produced by the full range of eye and head movements. The gaze can move in this space without affecting the hands and vice versa and these two spaces must be related in some way to support hand/eye correlation and coordination. This is indicated as another mapping. Movement of the torso affects both hands and eyes and the resulting disturbance effects must be similarly mapped onto the gaze space. **Figure 2** also indicates that memory will be necessary to record learning of significant and successful experiences, and we use a schema formalism for this.

Given this anatomy we can now define the initial state of the system prior to the experiment. The robot will be furnished with a framework upon which to learn hand-eye coordination and object interaction. This framework will support learning in the various sensor and motor modalities, coordination between modalities, and the creation and integration of schemas. Initially it will not contain any schemas or data on the coordination of sensor and motor systems, with this being learnt through exploration and interaction.

The goal of the experiment is for the robot to progress from the initial state to a state where it has control over its subsystems so that hand-eye coordinated reaching is achieved. We can measure attainment of the goal by an ability to reach for objects and to move them in the environment. A second objective is to achieve

¹The iCub is designed to be closely modeled on infant anatomy. For example, the eyes can saccade at speeds approaching human performance; we set the saccade velocity to be 80 degrees/s.

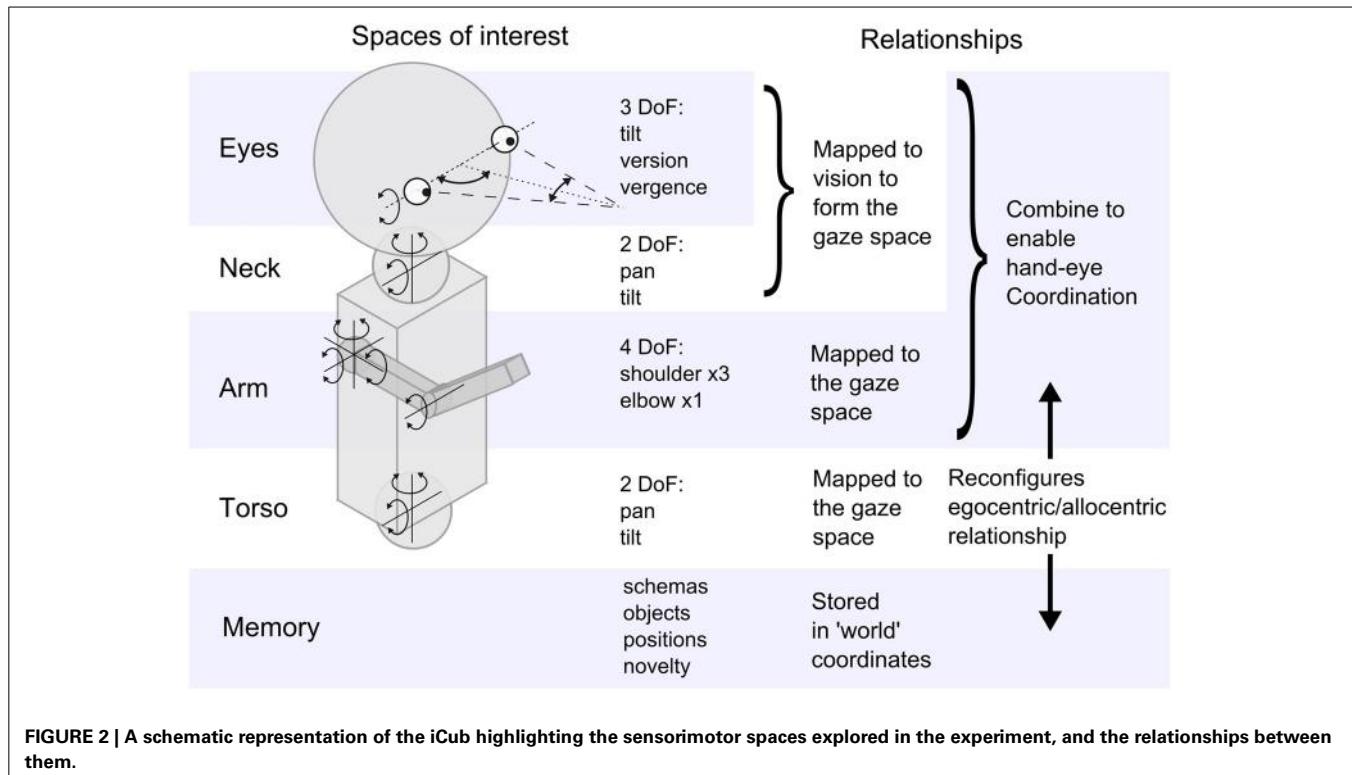


FIGURE 2 | A schematic representation of the iCub highlighting the sensorimotor spaces explored in the experiment, and the relationships between them.

this through a process of novelty-driven learning that models the development shown by infants. This means learning must not involve supervision, yet it must also be constrained within an acceptable rate for real robot systems.

2.1. DESIGN OF THE EXPERIMENTAL ARCHITECTURE—INTRINSIC MOTIVATION

There are several key concepts implemented in our experimental software that form the basis of our approach. These include a motivational mechanism, a staged learning framework, a spatial sensorimotor substrate, and a schema memory for the recall and generalization of previous experiences.

The first concern is intrinsic motivation: as our robot is not given any goals or tasks, how (or even why) can it perform actions? Extrinsic goals are not sufficient to explain all behavior—some behavior is essentially internally driven, and this is particularly significant for the developing infant. For example, in a quiescent state with no external demands or priorities there may be a range of possible actions available but no indication or experience of the outcomes of those actions. In such cases many robotic projects have used the idea of “motor babbling” to select the next action randomly, e.g., Caligiore et al. (2008) and Saegusa et al. (2009). This relates to an enactive view of cognition in which action is seen as part of sensory data gathering. An interesting line of investigation is the work of Kuniyoshi and Sangawa (2006), who study and simulate the class of general movements in the fetus as a bootstrapping stage for postnatal exploitation.

Following Bruner et al. (1976), we use novelty as a driver. Novel events that can be repeated and possibly correlated with other events are given high saliency. We prefer a broad and general

definition of novelty that can be widely scientifically applicable. So our mechanism for novelty is simply to define any new event as stimulating. This is very general in that it includes new external stimuli, new internal experiences (such as from muscles or proprioception), new forms of interaction, or new sequencing of known events. Whether an event is detected as new by the robot, depends on it being sufficient to be detected by the sensing abilities of the system, and whether or not it was predicted, i.e., had a prior representation of it being experienced before². For example, a visual stimulus will be detected as new because it appears in a new location, or has changed color (provided neither were predicted). A movement of the robot will be considered new if it results in a detectable position that has not been previously encountered. An action combination will be considered new if it results in a change of world state that was not predicted. As an event or perceived structure becomes familiar so it will no longer be novel and becomes of less interest. This means the scope of novelty will change and evolve: initially even basic movements of the body parts are novel but later on objects become more interesting, followed by interactions with moving objects, animated objects and people.

In our implementation we assign all distinct stimuli, objects or other sensed entities, with an excitation variable that is given a high value on first encounter. All excitations decay with time

²In our system visual detection is based on objects represented as patches of color. An object, consisting of a cluster of pixels of minimum radius 5 pixels, is considered to have changed if 20% of the pixels change. Prediction or expectation is determined by the sensory data matching, or partially matching, an existing stored schema.

and a habituation function provides a brief sensitization period followed by a decline in excitation on repeated stimuli. This excitation regime provides a saliency device and a winner-takes-all selector then acts as an attention mechanism. With this arrangement the focus of attention is attracted toward the items that have been the most novel most recently. Over time the decay function will cause past events to be forgotten, thus effecting a short-term memory, and it then becomes possible for an old experience to become stimulating again.

When attention is attracted to a novel stimulus then activity is initiated in the form of motor babbling. Sometimes the stimulus will have no further existence but sometimes an action may co-occur with a repeat of the stimulus. In our approach, a major assumption is based on repeated events; if a novel stimulus can be repeated when a given action is performed then the stimulus and the action are likely to be causally linked (Lee et al., 2007). Hence, repetition is an important part of motor babbling. When a babbling action apparently disturbs a sensory signal of interest then the system is strongly motivated to repeat the action, and if the effect is confirmed then an association can be recorded in the developing perceptual structures. This form of correlation is correspondence-based with tight temporal constraints for simultaneous events—in neuroscience the window for events that are perceived to be the same or connected is reported to be lower than 10 ms (Caporale and Dan, 2008; Markram et al., 2011). For actions that need to be completed before their effect can be correlated with a stimulus we note that the basal ganglia uses dopamine effects for identifying which of several ongoing actions is the correlating action (Redgrave et al., 2013).

Figure 3 shows the algorithm for this process. Motor action is driven by novel stimuli, with correlations (mappings) between sensor and motor pairings being reinforced through repetition. Global excitation is the summation of all the excitation variables and is used to select motor babbling when there has been a period with no novel events.

In our systems novelty usually comes in the form of an unexpected sensory stimulus (visual, tactile, audio, etc.) or a stimulus that correlates with a motor act (arm movement and proprioception, hand movement and visual regard, object contact and movement, etc.). In the former case the saliency mechanism uses excitation values to select the most novel stimuli to attend to. There is no threshold limit: the highest excitation wins. In the latter case, the algorithm compares sensor and motor pairs with those stored in memory. If the new event is not already stored, then it is deemed novel and selected for further exploration. If it is repeatable and temporally coincident, it becomes saved as reliable experience.

2.2. TASK LEARNING COMPLEXITY

The second design issue concerns the complexity of the task of learning how a many degree of freedom system is related and structured. This becomes very difficult and computationally expensive for high orders and for our 15 DoF robot it is impracticable to consider learning over all the motor systems at once. However, infants face an identical problem, and they solve it incrementally and in real-time. Infant development is characterized by the phenomenon of staged behavior, during which prominent sequences are readily observed, for example, sitting, crawling, and walking. Competence in a task is preceded by mastery of other subtasks, and such stages involve periods of learning followed by consolidation (Piaget and Cook, 1952). Transitions between stages are neither instantaneous nor absolute, as one pattern of behavior supersedes or merges into another as the underlying control schemas change (Guerin et al., 2013). Piaget's theories were extended by Kalnins and Bruner (1973) and Bruner (1990), suggesting mechanisms that could explain the relation of symbols to motor acts, especially concerning the manipulation of objects and the interpretation of observations.

Table 1 has been derived from the developmental literature and shows the sequence of development of motor control for

```

Begin action
If Global Excitation = low
    Motor-values := Select(random motor values);
    or (extract from any previously learnt {sensor, motor} pairing)
    goto Perform action

If Global Excitation = high
    Sensory-values := Select(stimuli with highest excitation)
    Motor-values := Retrieve from {sensor, motor} pair in mapping
    goto Perform action

Perform action
    Use Motor-values to perform a motor act
    Receive all resultant sensory inputs
    Search for novel states or values (previously undetected)
    If Novelty detected
        Set high excitation value for novel stimulus
        Record {sensor, motor} pair with initial weighting in mappings
    Else
        If retrieved action then increment Hebbian weight of stored {sensor, motor} pair
    Repeat (goto Begin action)
  
```

FIGURE 3 | Algorithm for novelty-driven action selection (derived from experiments in Law et al., 2011).

Table 1 | Infant development and learning targets.

Age (months)	Observed behavior	Robot targets
Pre-natal	Grasp reflex Butterworth and Harris, 1994	Grasp on tactile feedback
1	Sufficient muscle tone to support brief head movements Fiorentino, 1981	Constraint on head movement
1	Eyes and head move to targets Sheridan, 1973	Learning of saccade mappings
1	Saccades are few in number Maurer and Maurer, 1988	
2	More saccades Maurer and Maurer, 1988 and improved control Fiorentino, 1981	Refinement of saccade mappings
2	Head only contributes to larger gaze shifts due to lack of muscle tone Goodkin, 1980	Release of constraint on head motion, and beginnings of eye-head mapping
2	Involuntary grasp release Fiorentino, 1981	Release grasp when hand attention is low
3	Head contributes to small gaze shifts 25% of the time, and always to large gaze shifts Goodkin, 1980	Refinement of eye-head gaze control
3	Reach and miss Shirley, 1933 with some contacts Fiorentino, 1981	Reaching triggered by visual stimulation
3	Hand regard and hands to mouth Fiorentino, 1981	Initial learning of eye-hand mappings with return to "home" position
3	Clasps and unclasps hands Sheridan, 1973	Learning of raking grasp
4	Good eye and head control Fiorentino, 1981	Gaze mapping completed
4	Beginning thumb opposition Bayley, 1936	Enable independent thumb movement
5	Rotation in upper trunk Fiorentino, 1981	Begin torso mapping
5	Palmar grasp Fiorentino, 1981	Learning of palmar grasp
6	Successful reach and grasp Sheridan, 1973	Refinement of visually-guided reaching
7	Thumb opposition complete Bayley, 1936	Refined thumb use
8	Pincer grasp, bilateral, unilateral, transfer Fiorentino, 1981	Learning of pincer grasp
8	Crude voluntary release of objects Fiorentino, 1981	Voluntary release
9	Leans forward without losing balance Sheridan, 1973	Torso mapping complete

the period up to 9 months. It shows the cephalocaudal direction of development, beginning with the eyes and head, and flowing down through the arms, hands, and torso. Early grasping and ungrasping, appearing before birth and at 2 months, respectively, are reflexive, but are included here as they provide vital actions for the development of behaviors. They enable the infant to perform basic manual interaction, and thus gain additional sensory information, without having to wait for controlled grasping to appear. These early, reflexive, actions are likely to help bootstrap later behavior, and highlight the importance of the concept of staged development: that it significantly reduces the complexity of the learning task.

Table 2 shows a similar set of data specifically for the behavioral stages identified in the development of reaching. We note that early reaching is driven by tactile and proprioceptive feedback, before vision is well established. As vision improves, so too does the level of involvement of vision in the feedback process: early arm movements are triggered by visual stimuli; the first successful reaches are visually elicited, with the eyes fixated on the target and not the hand; later reaches use visual feedback to reduce the error between the hand position and the target. There is also an element of proximo-distal development, with control of the shoulder appearing before the elbow and hands.

Together with the infant behaviors are a suggested series of stages that a robot could follow to achieve the same performance. These have been generated by relating the infant data to the specification of the iCub robot. However, they are general enough as to be applicable to most humanoid robot platforms.

In the experiment described here we aim to reconstruct the first 5 months of development indicated by these tables as a series of behavioral stages on the iCub robot.

The phenomenon of staged growth has been linked to the existence of maturational or environmental constraints. Various forms of constraint can be identified that restrict the range of sensorimotor functionality available to the young infant. One example of underlying constraints is seen in the development of the newborn that proceeds in a cephalocaudal manner, with behaviors emerging sequentially down through the body and including looking, orienting, swiping, reaching, grasping, standing, and walking. We modeled these effects in our robot experiments by restricting the information and action possibilities available to the robot; thus the complexity of the learning space is reduced, with related restrictions on the behaviors produced. In particular, we focus on how maturational constraints and individual experience affect the emergence of stages. In our robotic systems constraints can be structured (Type A), or emergent (Type B). Type A constraints are analogous to maturation in neurological and physiological structures, and cover changes in myelination, sensory resolution, muscle tone, etc. In contrast, Type B constraints emerge from interactions and experience. As the infant/robot develops, both types of constraints can be released, through maturation or interaction, leading to new abilities and behaviors.

Type A constraints are considered to be hard constraints on the developmental trajectory due to the physical growth or maturity necessary for their removal. Individual infants develop at different

Table 2 | Reach development and learning targets.

Age (months)	Observed behavior	Robot targets
Pre-natal	Arm babbling in the womb De Vries et al., 1984	Proprioceptive-motor mapping of general movements
1	Hand-mouth movements Rochat, 1993	Learning of home position through tactile feedback
1	Directed (to the hemifield in which a target appears), but unsuccessful, hand movements von Hofsten and Rönnqvist, 1993; Ennouri and Bloch, 1996	Initial mapping of general movements to vision
1	Initial reaching is goal directed, and triggered by a visual stimulus, but visual feedback is not used to correct movements mid-reach Bremner, 1994, p. 38	Visual stimuli trigger general reach movements
3	Infants often move their hand to a pre-reaching position near the head before starting a reach Berthier et al., 1999, which then follows the line of sight Bruner, 1968, p. 44	Reaches conducted from "home" position
3	Infants engaged in early reaching maintained a constant hand-body distance by locking the elbow, and instead used torso movements to alter the distance to targets Berthier et al., 1999	Constraints on elbow movements reduce learning space
3	Successful reaching appears around 3–4 months after birth Shirley, 1933; Fiorentino, 1981; Berthier et al., 1999; Berthier and Keen, 2006	Primitive hand-eye mapping
3	Gaze still focused on the target and not the hand Clifton et al., 1993; Butterworth and Harris, 1994; Clifton et al., 1994; Berthier and Carrico, 2010	Reaches are visually elicited, but without continuous feedback
4	From 4 months, infants begin to use visual feedback to refine the movement of the hand White et al., 1964	Begin to map joint-visual changes and use visual feedback to correct reaches
4	As infants age their reaching becomes straighter, with the hand following the shortest path Carvalho et al., 2007	Refined reaching with smooth and direct movements

rates, making timings of constraint releases difficult to define, however, the trajectories tend to be similar, following a regular sequence of stages. A timeline is presented in Law et al. (2011) and can be applied to a developing robot by using internal state variables as the indicator to trigger removal of constraints in a semi-structured manner (Lee et al., 2007). This will cause the robot to follow the general infant trajectory, where the timings of constraint release are based on its own individual circumstances.

Type B constraints are caused by external factors that effect development, such as the level of stimulation in the environment or the amount and form of interaction with carers. The strong influence of these factors on the order in which development occurs has been recorded in observation and demonstrated in various experiments. For example the use of a "sticky mitten" to compensate for the lack of competence in grasping, facilitated infants with a precocious and greater level of manual interaction with objects (Needham et al., 2002).

Both types of constraint play an important role in this experiment. The development of muscle tone, a Type A constraint, is cited as a driver for cephalocaudal development, and provides us with our basis for creating the pattern of behavior in **Table 1**. As we are not able to accurately model this type of development, we simulate it as a series of constraints preventing movement at each set of joints. These constraints are released in sequence, starting with the eyes at the outset, and progressing down the body as the experiment continues. Whereas muscle tone is likely to be related to age, our constraints are related to level of ability, as our developmental sequence has a much shorter time scale than that of the infant.

Other Type A constraints, in the form of sensory availability and resolution, are used to shape reaching actions. Initial arm movements are formed using tactile and proprioceptive feedback without any visual information. Once vision is active, it can be incorporated into reach learning, but resolution in the infant gradually increases, and we model this growth. Early visually triggered reaches generate very coarse visual stimuli, so result in inaccurate swiping behaviors in the general direction of objects. As vision and gaze control improve, so does the quality of reach. However, visual feedback is not enabled to guide reaching until visually elicited reaches have become successful. Due to the requirement for physical interaction during reach learning and the need to avoid potentially harmful robot actions, we conduct these stages in simulation and transfer them to the robot when accurate reaching has been achieved.

In addition to these Type A constraints, our experiment relies on Type B constraints arising from the environment. Although the effects are often quite subtle they can also be quite pronounced, for instance the number and positioning of stimuli impact on the extent of learning. Due to the size and nature of the experiment these influences make it very difficult to measure their effects or replicate data precisely. To show the impact of these constraints we investigate how changing the environment affects learning of gaze control. Details of the constraints used in this experiment are given in **Tables 3, 4**. In **Table 3**, the Type A constraint on the torso and arm learning is the same, i.e., restriction of movement due to immaturity. However, if these two components tried to learn in parallel then a number of variables and unconstrained degrees of freedom would be active at

Table 3 | Constraints used to structure behavioral stages on the robot.

Constraint	Type	Effect	Removal trigger
Environment	B	Affects data available for learning at all stages	None. Influenced by robot and experimenter
Eye motor	A	Prevents eye motion	Start of experiment
Neck motor	A	Prevents head motion	Threshold on eye control
Neck learning	B	Neck learning requires accurate eye control	Emerges as eye control develops
Shoulder motors	A + B	Prevents arm movement	Threshold on gaze control, exclusive of torso learning
Elbow motors	A + B	Limits forearm extension/flexion	Threshold on gaze control, exclusive of torso learning
Reflex grasp	A	Causes hand to close on tactile stimuli	Active until reaching threshold attained
Controlled grasp	A	Prevents voluntary grasping of objects	Released with shoulder
Torso motor	A + B	Prevents motion at waist	Threshold on gaze control, exclusive of arm learning

Table 4 | Constraints used to structure reaching stages in simulation.

Constraint	Type	Effect	Removal trigger
No vision	A	Arm movements learnt through tactile and proprioceptive feedback only	Start of experiment
Crude gaze fields (large)	A	Arm movements coarsely correlated with vision	Threshold on maturity of internal structures
Fine gaze fields (small)	A	Fine correlation between hand position and vision	Threshold on development of reaching
Visual feedback	A	Prevents visual guidance during reaching	Threshold on successful reaching

the same time. It would be very difficult to identify which motor movements caused which effects, making it very difficult to learn anything meaningful. As a result, the constraints are used to prevent them learning at exactly the same time, but the order in which they learn is flexible being based on stimulation and events in the environment. Consequently, we label these as containing both type A and B constraints. It is equally possible that the two could alternate in their learning, with constraints being intermittently applied to alternating components. Neck learning is also shown in **Table 3** as Type B; this is because neck learning does not need a threshold or trigger as it is only effective when eye control is well developed. Hence, it can be permanently enabled but will only emerge as and when the eye system achieves sufficient competence. Such emergence is typical of Type B constraints.

2.3. MAPPING TECHNIQUE

The third key issue concerns the design of a suitable computational substrate that will support the representation of whatever sensorimotor structure is discovered by experience. This involves spatial data as can be seen from the robot hardware in **Figure 2**. This figure suggests the fundamental spaces produced by the sensory-motor configuration of the iCub robot and, following the embodiment principle, this will vary for different anatomies. Considering the staged organization mentioned above, we designed the architecture shown in **Figure 4** to capture the relations and mappings indicated in **Figure 2**.

Learning data in this experiment is based on visual and proprioceptive data. That is, the image data collected by the two cameras, and the information from the position of each joint. Tactile sensors trigger reflexive grasping, but are not directly used in learning. The main components of the architecture are as follows:

- Visual stimuli on the camera sensor are encoded on a 2D retinotopic map and linked to 2D motor maps for the eyes and neck. These enable the robot to learn the correspondence between moving the eyes and neck, and movements of visual stimuli. A mechanism for gaze control based on biological data interacts with both the eye and neck motor maps to generate stereotypical gaze shifts. The combination of both eyes and neck displacements defines the gaze space—the 3D egocentric model of space used to coordinate the robot's actions (see Law et al., 2013, for further details).
- 4D arm motor movements are mapped to a portion of the gaze space, for hand-eye coordination.
- 2D torso motor movements are mapped to the gaze space to define how body movements affect the movement of visual targets.
- The memory schema records the positions and details of objects in the 3D gaze space, but the relationship between the current gaze direction and the remembered positions changes as the robot moves. Data from the learnt torso mapping is used to transform remembered positions into relative gaze positions.

The architecture is thus a cross-modal representation of the robot and its personal space. At the core of our architecture is the 3D egocentric gaze space, which maps the proprioceptively-sensed gaze direction of the eye and head to the visual space of the retina and the proprioceptively-sensed position of the limbs in joint space. This building up of an internal body model from a collection of smaller spaces has been investigated by others, e.g., Morasso and Sanguineti (1995) and Fuke et al. (2009) but the key challenge is in keeping the computational demands of the techniques within the bounds of biological plausibility.

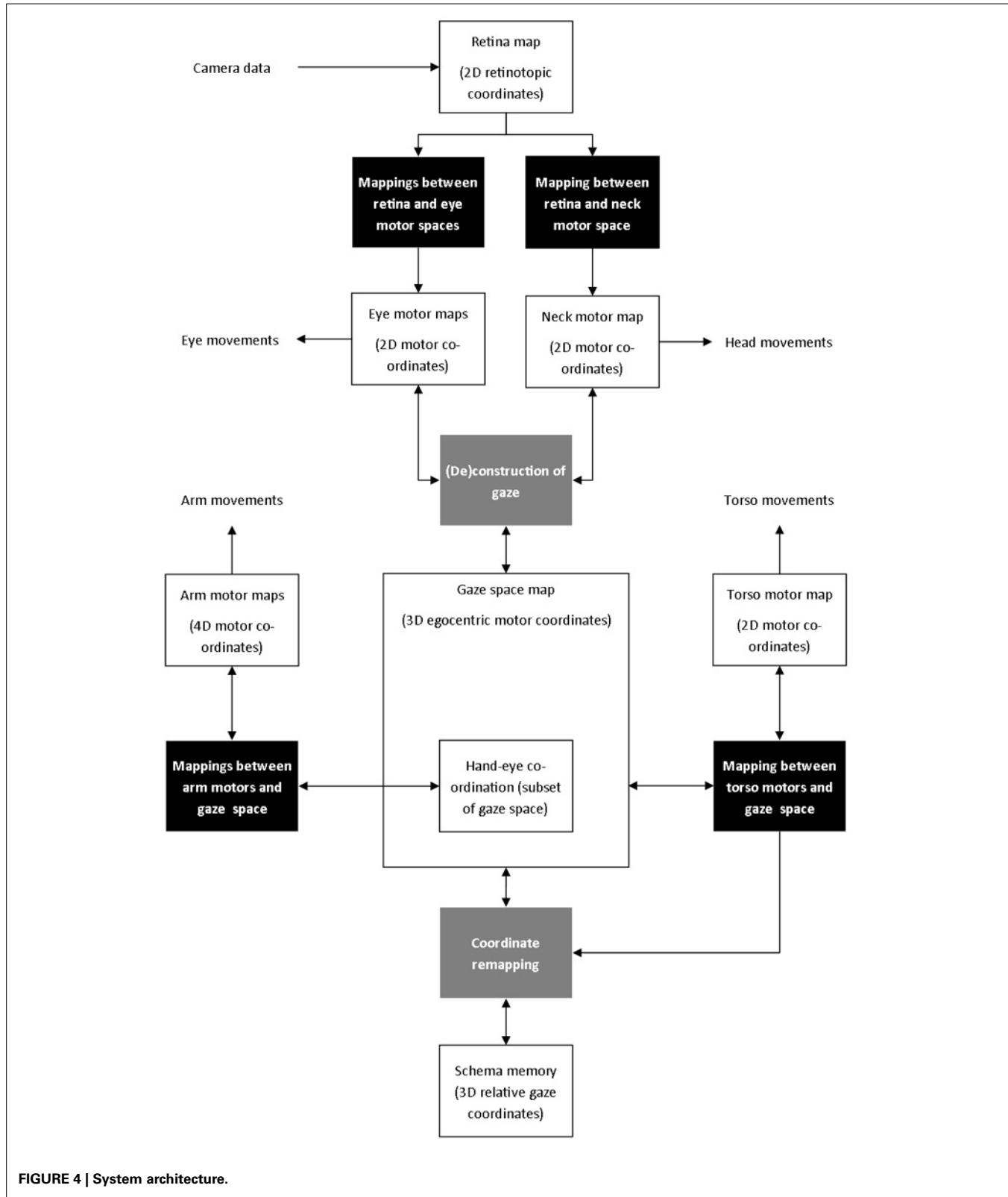


FIGURE 4 | System architecture.

Piaget suggested infants first construct an egocentric representation of space through sensorimotor interaction, and that this gradually gave way, over the first year of life, to the ability to locate objects in relation to external landmarks (Piaget and Inhelder, 1956). More recently, this has given way to the idea of infants developing an allocentric representation of space, based on a variety of coding mechanisms (Newcombe and Huttenlocher, 2006). This shift is most noticeable in the latter half of the first

year (Acredolo and Evans, 1980; Newcombe and Huttenlocher, 2006; Vasilyeva and Lourenco, 2012), beyond the period of our current investigation, but has also been suggested to appear as early as 4 months (Kaufman et al., 2006; Bremner et al., 2008). The shift from egocentric to allocentric representation is noted to be slow, and could be related to a number of factors including identification of visual landmarks, rotation of the torso, and crawling (Newcombe and Huttenlocher, 2006; Vasilyeva and Lourenco, 2012), and that it could be impaired by cognitive load (Kaufman et al., 2006). Our vision system is not capable of identifying relationships between objects, nor does the robot perform any relocation of the body until the torso develops late in the experiment. Therefore, we currently restrict our model to the early egocentric and proprioceptive representation of space.

Motor babbling generates candidate data for learning this sensorimotor coordination. The discovered associations between stimuli properties and motor acts represent important information that will support further competencies. For example, in controlling the eyeball to move to fixate on a target it is necessary to know the relationship between the target distance from the center of the retina and the strength of the motor signals required to move the eye to this point. As targets vary their location in the retinal periphery so the required motor command also varies.

We use a mapping method as a framework for sensorimotor coordination (Lee et al., 2007). A *mapping* consists of two 2D arrays (or *maps*), representing sensory or motor variables, connected together by a set of explicit links that join points or small regions, known as *fields*, in each array. Although three dimensions might seem appropriate for representing spatial events, we take inspiration from neuroscience, which shows that most areas of the brain are organized in topographical two-dimensional layers³ (Mallot et al., 1990; Braitenberg and Schüz, 1991). This remarkable structural consistency suggests some potential advantage or efficacy in such two-dimensional arrangements (Kaas, 1997).

Fields are analogous to receptive fields in the brain, and identify regions of equivalence. Any stimulus falling within a field produces an output. A single stimulus may activate a number of fields if it occurs in an area of overlap between fields. Further studies of the map structure and how it relates to neural sheets in the brain is presented in Earland et al. (2014).

For the saccade example, a 2D map of the retina is connected to a 2D array of motor values corresponding to the two degrees of freedom provided by the two axes of movement of the eyeball (pan and tilt). The connections (representing the mapping) between the two arrays are established from sensory-motor pairs that are produced during learning. Eventually the maps are fully populated and linked, but even before then they can be used to drive saccades if entries have been created for the current target location.

Mappings provide us with a method of connecting multiple sensor and motor systems that are directly related. This is sufficient for simple control of independent motor systems, such

as by generating eye-motor commands to fixate on a particular stimulus. However, more complex and interdependent combinations of sensor and motor systems require additional circuitry and mechanisms in order to provide the required functionality. For example, audio-visual localization requires the correlation of audible stimuli in head-centered coordinates, with visual stimuli in eye-centered coordinates. The system has the added complexity that the eye is free to rotate within the head, making a direct mapping between audio and visual stimuli impossible. Just as in the brain, careful organization and structuring of these mechanisms and mappings is required.

To control coupled sensorimotor systems, such as the eye and head during gaze shifts, we take inspiration from the relevant biological literature (Guitton and Volle, 1987; Goossens and van Opstal, 1997; Girard and Berthoz, 2005; Freedman, 2008; Gandhi and Katnani, 2011). Our aim is to reproduce the mechanisms at a functional level, and connect them to form an appropriate abstraction of their biological counterparts. We do not endeavor to create accurate neurophysiological models but rather to create plausible models at a functional level based on well-established hypotheses. Consequently we use mappings to transform between sensorimotor domains, and incorporate standard robotic sensors and actuators, and low-level motor control techniques in place of their biological equivalents.

2.4. SENSORIMOTOR MEMORY

The final design issue concerns the requirement to remember learned skills. The system as described above has a rich sensorimotor model of the immediate events being experienced but has limited memory of these experiences.

Until this point the mappings have acted as the sole memory component, storing both short term sensory, and long term coordination information. The sensory events are mainly spatially encoded, in the robot's "egosphere" as indicated above, and these have short term memory—when their excitation decays they may be experienced again as "new" events. On the other hand, the coordinations between motor and sensory subsystems are stored as connections and thus represent long term memories (with scope for plastic variation). These are also mainly spatially encoded experiences and so represent, for example, how to reach and touch an object seen at a specific location. What is not represented is any sensorimotor experience that has temporally dependent aspects. For example, consider a sequence of actions such as: reach to object, grasp object, move to another location, release object. This can be seen as a single compound action (move object) consisting of four temporally ordered actions.

For this reason we introduced a long term associative memory mechanism that supports: the memory of successful basic action patterns; the associative matching of current sensorimotor experience to the stored action patterns; the generalization of several similar experiences into single parameterized patterns; and the composition of action chains for the temporal execution of more powerful action sequences. A concomitant feature of such requirements is that the patterns in long term memory should be useful as predictors of action outcome—a function that is unavailable without action memory. Inspired by Piaget's notions of schemas (Piaget and Cook, 1952) we implemented a schema

³Each field can hold a range of variables, effectively providing a 2.5D representation. This is how space is represented with depth as a field value.

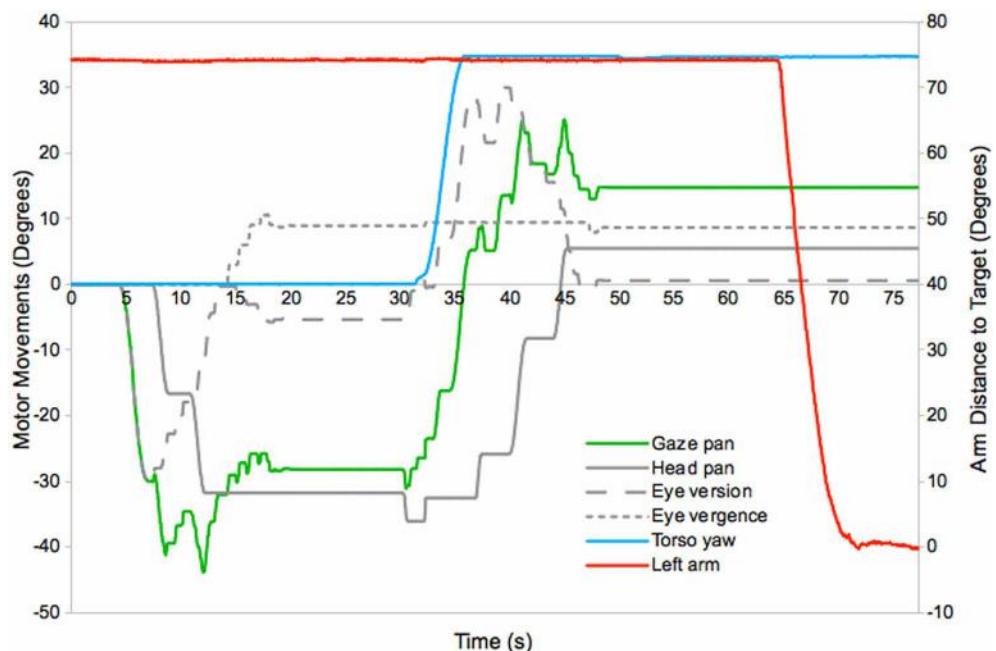


FIGURE 5 | Motor dynamics in the horizontal plane during a typical gaze and reach action (see text for details)⁴.

memory system that stores action representations as triples consisting of: the pre-conditions that existed before the action; the action performed; and the resulting post-conditions (see Sheldon, 2012, for further details). The schema memory provides long term memory in order to prevent repeated attention on past stimuli, and to match previous actions to new events. This formalism has been used by others, e.g., (Guerin et al., 2013), and is a flexible and general representation that allows extensions and supports all the above requirements.

3. RESULTS

There are four significant results from this longitudinal experiment.

3.1. EMERGENCE OF DEVELOPMENTAL STAGES

The first result concerns the emergence of a series of distinct qualitative stages in the robot's behavior over the duration of the experiment. The results described here used maturity levels for constraint release that previous experiments suggested as reasonable. This gave competence for reaching to be performed with an end-point accuracy of 1 cm, which is sufficient for the iCub to grasp 6–8 cm objects. Further data on the effects of staged constraint release can be found in Shaw et al. (2014).

An example of the motor dynamics exhibited during reaching, using maps learnt in this experiment, is given in **Figure 5**. This shows the use of the eyes, head, torso, and arm joints to gaze to a novel target, bring it into reach, and place the hand at its location. At around 4 s into the experiment a novel target appears and the robot initiates a gaze shift. This is produced using the eye and head motor movements mapped to the location of the stimulus in the retina map. The eye is the first system to move, and fixates

on the target at around 6 s. The head then begins to contribute to the gaze shift, and the eye counter-rotates to keep the target fixated (the mapping resolution and dynamics of the system result in jerky head movements and some fluctuation of the gaze direction). The gaze shift completes at 14 s and is followed by a separate vergence movement to determine the distance to the target (this has a small effect on the gaze direction, which is based on readings from the dominant eye). Full fixation occurs around 19 s. Next the robot selects a torso movement to position the target within the reach space. This takes place between 31 and 35 s and is accompanied by compensatory eye and head movements, which complete at 48 s. Finally arm movements are triggered at 65 s, which result in the hand arriving at the target at 71 s.

Table 5 records these stages and also their rapid rate of development. As previously explained, the early reaching actions are first learnt in simulation before being integrated with other actions learnt on the real robot. Aside from this deviation for safety purposes, all actions are learnt on-line on the robot. All actions learnt in simulation are performed at a speed equivalent to real-time on the physical robot in order that the results are comparable⁵.

Tables 6, 7 expand on the detail and show the time point when each stage was observed to first appear.⁶ The resultant behavior patterns are similar to those in **Tables 1, 2**, with the omission

⁴The version angle is the combined pan angle of the two cameras.

⁵A time lapse video of the longitudinal process on the iCub can be found at <http://youtu.be/OhWeKIyNcj8>

⁶Videos of the robot performing some of these stages, with basic reaching and torso movements, can be found at http://youtu.be/_ZIkB8FZbU and <http://youtu.be/3zb88qYmxMw>. Full reaching and torso control is shown in <http://youtu.be/OhWeKIyNcj8>

Table 5 | Observable experimental behaviors.

Behavior	Description	Duration (min)	Platform
Fetal babbling	General arm movements	10	Simulator
Saccading	Eye movements only, trying to fixate on stimuli	20	Robot
Gazing	Eyes and head move to fixate on stimuli	40	Robot
Swiping	Arms make swiping actions in the general direction of visual stimuli	10	Simulator
Visually elicited reaching	Reaches toward visual targets with some success	10	Simulator
Guided reaching	Successful and smoother reaches toward visual targets	60	Both
Torso movement	Moves at waist to reach objects	20	Robot
Object play	Grasps objects and moves them around	40	Robot

Table 6 | Behaviors observed on the iCub.

Behavior	Description	Time of appearance (min)
Saccading	Eye movements only, trying to fixate on stimuli	0
Gazing	Eyes and head move to fixate on stimuli	20
Guided reaching	Successful and smoother reaches toward visual targets	60
Torso movement	Moves at waist to reach objects	120
Repeated touching	Repeatedly reaches out and touches objects	140
Pointing	Points to objects out of reach	160
Object play	Explores object affordances and actions	170
Stacks objects	Places one object on top of another	210
Learning ends	Experiment ends	230

Table 7 | Behaviors observed in simulation.

Behavior	Description	Time of appearance (min)
Fetal babbling	General arm movements	0
Pre-reaching position	Moves hand to the side of the head before reaching	10
Swiping	Arms make swiping actions in the general direction of visual stimuli	10
Visually elicited reaching	Reaches toward visual targets with some success	20
Guided reaching	Successful and smoother reaches toward visual targets	30
Learning ends	Refined hand-eye coordination	90

of some of the finer details. In general, however, robot development progresses along cephalocaudal and proximo-distal learning directions. Whilst this is to be expected due to the choice of constraints, the experiment also demonstrates the efficiency in

this learning pattern. **Tables 6, 7** show the time taken for the robot to advance from the experiment's initial state to the final goal state is less than 4 h. This is possible because the constraints limit the size of the learning space, whilst the resultant ordering of stages generates a sequence whereby earlier behaviors create data for bootstrapping learning of later behaviors. For example, eye saccades provide data for learning of gaze control, which is in turn used as a basis for hand-eye coordination. Similarly, in the arm system, the staged increase of gaze field resolution enables mappings to be created that are initially very sparse, but which are then refined as resolution improves. Without bootstrapping, the high dimensionality of the space means considerably more learning will be required to reach a similar level of ability across all areas. For further material on this, for head and eye learning see Shaw et al. (2012), and for reaching see Law et al. (2014).

Table 8 highlights the dimensionality issue, and shows how stages break down the mappings into manageable chunks. 15 degrees of freedom in the motor space are mapped to 15 dimensions in the sensory space, using seven core stages. Movements of the eyes, head, and torso are all mapped to the gaze space to provide visual orientation, but the series nature of the joints requires learning to follow the pattern eyes-head-torso. Reaching is learnt through four stages, with both arms learning in parallel. The four stages correspond to a shift from tactile and proprioceptive to visual mapping, followed by improvements in visual resolution.

3.2. IMPACT OF CONSTRAINTS

The second result concerns the impact of different constraints, and the timing of their removal, on learning. We use the eye and head components of the gaze system to illustrate the effects of both Type A and Type B constraints on the development of gaze control.

Gaze control is learnt in two stages: mapping of visual changes to the eye motors, and mapping of visual changes to the neck motors. In reflection of the human gaze system, a stabilizing ocular reflex causes the eyes to rotate to compensate for movements of the head, and maintain fixation on a stimulus. This prevents a direct mapping from neck motors to vision, as the eye reflex minimizes visual change. Therefore, the mapping must take into account changes in eye position and their known effect on visual stimuli, which requires a well developed eye mapping [for a detailed description of the gaze-learning algorithm see Law et al. (2013)].

We simulated the documented effect of poor muscle tone in the neck by imposing a constraint on head movement. We varied the time at which this constraint was released to model a Type A constraint, and varied the level of stimulation in the environment to model the effect of a Type B constraint. In the case of the Type A constraint, we compared the effect of reducing the head constraint at 10 min intervals over seven 1 h learning periods. In most cases this resulted in the eye and head systems learning *in parallel* for part of the learning period. In the case where the constraint was removed at time $t = 0$, an emergent constraint appeared with the head system failing to learn correct movements until the eye mapping was partially developed. Over the whole course of the learning period, this resulted in slower learning as both systems attempted to develop in parallel. In comparison, when the constraint lifting was delayed, the eye mapping was initially able to develop more rapidly on its own. When the constraint was eventually lifted, data from the eye map was available to support head learning, resulting in immediate learning of correct movements.

Figure 6 shows the number of links in the head mapping learnt over time. The most links were learnt when the constraint was released between 10 and 20 min after eye learning had begun. This

represents a trade-off between the level of eye control required to support head learning, and the remaining time available for learning.

In order to evaluate the impact of a Type B constraint on the eye and head development, we varied the level of stimulation within the environment. Previously a selection of static visual stimuli had been available for the robot to select as targets for saccade learning, however, in this case only a single static target was presented centrally in front of the robot. The effect of this was to limit the size of eye motor movements that could be made without losing sight of the target, thus limiting eye learning. Here, removal of the constraint on head movement enables the target to appear off-center of the eye, simulating its appearance at new locations.

Figure 7 shows the coverage in the eye map, in terms of fields, as links are learnt. This is a measure of how much of the visual space can be reached by a known saccade. With only a single, stationary visual stimulus available, eye learning saturates at around 50% coverage. The effect of repositioning the stimulus can be clearly seen in the periods following the constraint removal, where coverage increases to around 80% without saturating. Further explanation of this phenomena can be found in Shaw et al. (2012).

These results show that both types of constraint impact on learning in significant ways. Maturational constraints prevent specific abilities, and limit the size of the learning space, whereas environmental constraints limit the complexity of the stimuli, and result in emergence of behaviors. Our experience is that both are required to drive efficient learning, but a balance is required. Too little constraint results in over stimulation, and problems in identifying correspondence, whereas over-constraint restricts and slows learning. Other mechanisms for releasing constraints are possible, e.g., Nagai et al. (2006) who compare an error measure method against fixed time scheduling.

Table 8 | Learning times using developmental processes.

System	Motor DoF	Sensory map dimensionality	Stages	Learning time (min)
Eyes	3	3	1	20
Head	2	2	1	40
Tactile		1		
Torso	2	3	1	20
Arms	4*2	3*2	4	90

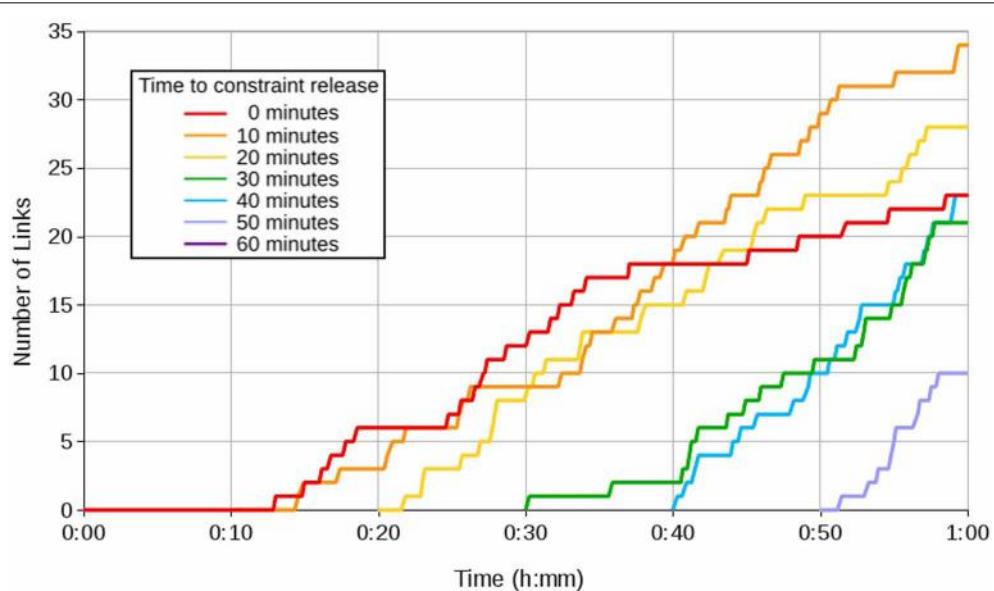


FIGURE 6 | Graph showing head learning with a Type A constraint lifted at 10 min intervals.

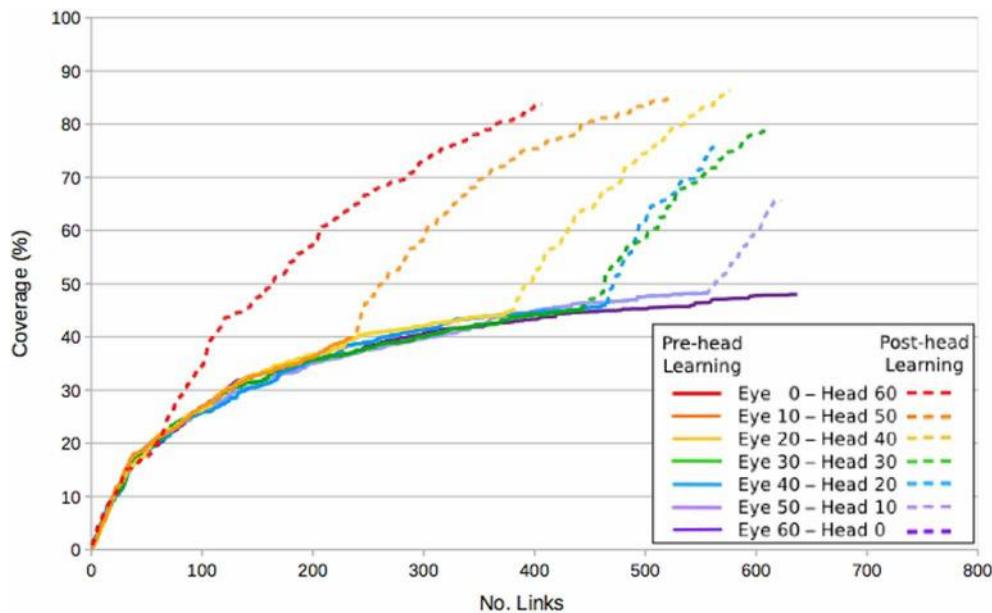


FIGURE 7 | Graph showing the effect of a Type B constraint on eye learning, when the head constraint is lifted at 10 min intervals.

3.3. IMPACT OF BOOTSTRAPPING BETWEEN STAGES

The third result concerns the impact of bootstrapping between stages, that is, the value of priming new behavior with previously learnt data. We use the problem of learning arm reaching to illustrate this.

In order for the robot to be able to reach to and pick up objects, it is very desirable that the trajectory of the hand follows a reasonably direct route to the target, avoiding dangerous configurations, obstacles, and possible damage to the robot. To achieve this we developed a vector based reaching algorithm using an adaptation of our mapping technique. For each arm a 4-to-3 dimensional mapping is created between the joint space of the shoulders and elbow, and the gaze space. This enables learning the correspondence between arm postures and the corresponding position of the hand in the visual space. An important addition is the ability of fields in both maps to store vectors. These allow movement directions in the gaze space to be mapped to motor movements in the joint space by performing and observing small movements of the arm. As vectors are stored as part of the field data, movements are learnt in correspondence to particular arm poses. When reaching, combinations of vectors from the current or nearby postures can be used to move the hand in a desired direction.

Although it is possible to learn this mapping in one stage, using our novelty-driven motor babbling, we have found that learning can be made much more efficient by using multiple stages, and using data learnt in earlier stages to bootstrap learning in later ones.

We note that the eyes of the fetus do not open until 26 weeks after conception, and that any vision is likely to be very limited. However, arm movements and tactile perception appear at 7–9 weeks, and there is the possibility for early learning through proprioception and tactile feedback. To simulate this we created a

very basic model of activity in the womb, through which simple arm movements are learnt using coarse proprioception. These are generated by motor babbling, and learning is triggered by tactile stimulation resulting from interaction with a modeled uterine wall. After 10 min of learning, a range of proprioceptive arm positions have been generated corresponding to these interactions, without any information on their position in space being stored. This data is then used to bootstrap hand-eye coordination and reaching.

In our experiment, we consider how even very primitive bootstrapping is important. During the immediate post-swiping stages in **Table 5** the robot performs hand regard. That is, it looks to the position of the hand and makes small movements in several directions to generate the vector mapping described above. As the vectors are only valid for the pose in which they are learnt, hand regard must be performed over a range of poses for them to be useful to control reaching. The bootstrapping data from the previous stages provides a set of known positions at which hand regard can be performed and, due to the ballistic character of much of the motor babbling behaviors, the locations tend to be at the extremities of the operating (reachable) space. This distribution in space is an advantage because movements between the locations provide a good covering of the space, whereas without this data, hand regard would tend to cluster around the central area and take much longer to explore the extremities. **Figure 8** shows images of the arm fields generated after 10 min of hand regard and reach learning. Using bootstrapping produces 36 fields with an average of 8.4 vectors per field, while without bootstrapping there are 22 fields with an average of 15.9 vectors per field. This shows how learning without bootstrapping is centered around a smaller set of configurations. Further data on the stages of reach learning can be found in Law et al. (2014).

Another aspect of staged transfer is seen in the removal of the gaze field constraint shown in **Table 4**. When vision is first used the generated fields are restricted to a large radius (0.7) and hence the covering of space is coarse. After 15 fields are produced the field size constraint is lifted (to radii = 0.2) and then the number of fields increase to 33 before the next stage. Thus the spatial covering becomes more exact and more accurate movements can be made. This differentiation of coarse or diffuse values into finer resolutions is seen in other developmental studies, e.g., regarding visual immaturity, Nagai et al. (2011) have shown how early sensorimotor associations formed during periods of poor discrimination can continue to be important when much finer discrimination has been achieved.

The results in **Figure 9** show the distances covered by the hand when reaching to a set of predefined targets using learnt

vector-based reach mappings. The three different data sets correspond to three different learning strategies: the first uses the method described above, performing hand regard at the poses in the bootstrapping data. The second ignores the bootstrapping data, but performs hand regard at positions it encounters whilst trying to reach to a target. The third uses neither bootstrapping nor hand regard, and only learns vectors corresponding to movements it makes whilst trying to reach to the targets. The main difference between these last two is that with hand regard the robot learns vectors in multiple directions, whereas without hand regard it can only learn vectors corresponding to the direction of motion. If no suitable vector to direct reaching was known, then a random movement is made. Learning was conducted for the same duration for each approach and then the mappings used to control reaches to 24 target positions, 12 for each arm, distributed throughout the robot's reachable space.

The results display the clear advantage in using bootstrapping data from previous stages. **Table 9** illustrates this by using deviation from the most direct path as an error measure. By comparing the average distance covered by the hand to the ideal straight-line path, of the three approaches, the one using bootstrapping resulted in a near ideal case.

3.4. LONG-TERM MEMORY FOR IDENTIFYING NOVEL EVENTS

Our fourth significant result shows how a memory of experienced actions enables appropriate responses to novel events and thus provides a framework for the emergence of new action skills.

As described in section 2.4, without a long-term memory the sensorimotor mappings can only support repeated actions

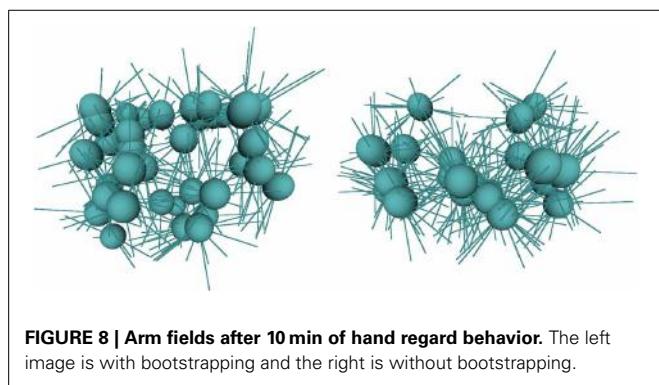


FIGURE 8 | Arm fields after 10 min of hand regard behavior. The left image is with bootstrapping and the right is without bootstrapping.

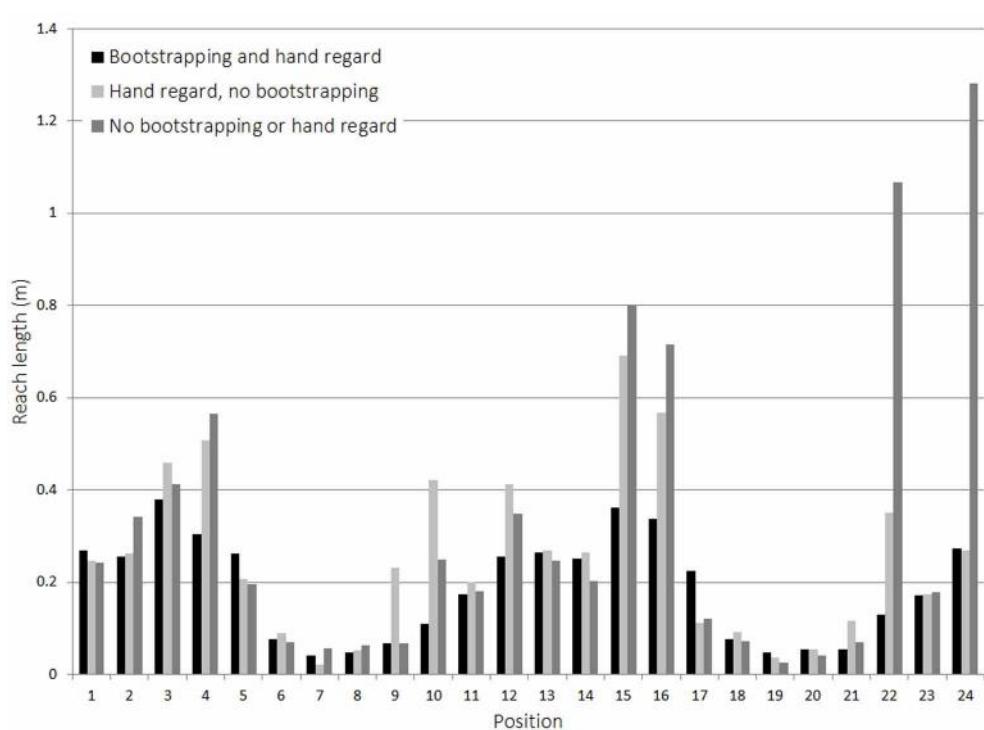


FIGURE 9 | Effect of bootstrapping on reach learning.

Table 9 | Reach length comparison.

Learning method	Average hand trajectory length compared to the direct path (%)
Bootstrapping and hand regard	107.5
Hand regard, no bootstrapping	149.5
No bootstrapping or hand regard	179.5

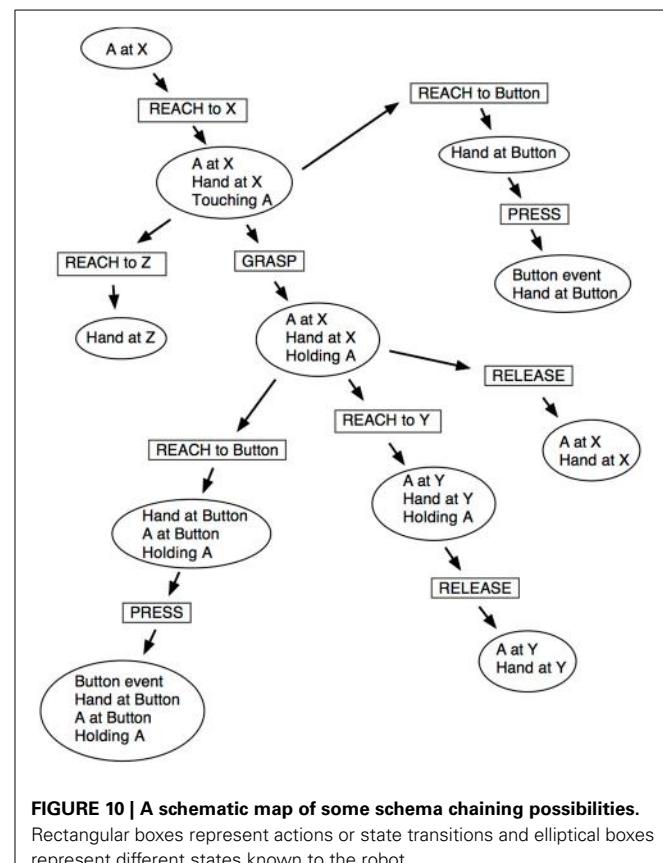
over short-term events, hence we use a schema learning mechanism that can record more complex actions and their consequences (Sheldon and Lee, 2011). A schema is a structure that encodes the context in which an action may be performed together with the action detail and its result, or more formally: <preconditions : action : postconditions>. An example schema for touching an object, A, at a specific location can be written:

<A at (35, 66) : Reach to (35, 66) : A at (35, 66), Hand at (35, 66), Touching A>

Schemas are created when an action is first performed, using the action and a set of observations. They also carry excitation values and data relating to the probability of their occurrence. Schema recall occurs when a new sensory event is detected and the schemas are scanned to find those that most match the current situation. This is achieved by exciting the schemas using a combination of the novelty of the current experience and their similarity to past experiences. The level of excitation increases with the novelty of the current sensation and the similarity to a remembered sensation (see Sheldon, 2012, for full details of the schema creation, matching, and generalization algorithms).

Just like stimuli, schema excitation values decrease as they are used. This means newly discovered schemas are more likely to be repeated and tested. Schema probability values track the likelihood of the schema succeeding and the more excited and predictable schemas are selected in preference to less excited and unpredictable ones. The result is that, initially, simple but reliable schemas are selected and explored. However, as their excitation levels drop more complex and potentially useful behaviors come to the fore. This promotes exploration when there are few immediate novelties, and can result in unexpected behavior. For example, in the later stages of the experiment the iCub had learnt schemas for reaching, pressing buttons and grasping objects. The iCub next learned that it could reach to, and grasp, an object, and that it could move that object by reaching to new locations. **Figure 10** illustrates how these actions can occur. The diagram indicates some possible states of the sensory data that schema actions can cause to change. At the top-left in **Figure 10** an object, A, is known to be located at position X. The sequence of schema applications along the diagonal toward bottom-right correspond to grasping an object and moving it to another place. Several other schemas are illustrated: the earlier schema of grasping but not moving an object is at center-right; and the pressing action is shown at top-right.

After the reaching-whilst-holding schema had become established, the iCub discovered that it could conduct a pressing action whilst holding an object; this was composed from the two above

**FIGURE 10 | A schematic map of some schema chaining possibilities.**

Rectangular boxes represent actions or state transitions and elliptical boxes represent different states known to the robot.

unrelated actions and is seen at left-bottom in **Figure 10**⁷. The motivational conditions that caused this, through the excitation and matching of prior experience to new situations, demonstrated how two unrelated actions may be combined to form new skills, and opens up the exciting prospect of learning tool use.

An important property of the schema framework is the ability to make generalizations. The generalization mechanism produces schemas containing parameters which can be populated based upon the current experiences of the robot when being executed. Beyond simply determining which aspects of the schema may be interchangeable with other values as many existing schema systems do, this mechanism attempts to find generalizable relationships between the preconditions, the action and the post-conditions of a schema. This allows the generalized schemas which are produced as a result of this process, to represent the agent's hypotheses about how an interaction may work at a more abstract level (see Sheldon, 2012, for further details).

Along with stages, generalization offers the means to reduce complexity in the learning environment. **Table 10** shows the number of schemas learnt to enable the robot to be able to reach and touch objects at any location within its workspace, or point to those out of reach. Without stages all combinations of stimuli and events create potential schemas, and so the prohibitive numbers of robot actions mean simulation is necessary to investigate

⁷A video of this behavior can be found at <http://youtu.be/VmFOobKd9A>

this aspect. These results show that staged learning dramatically reduces the number of schemas that are learnt by a simulated robot, and that combining a staged approach with the generalization mechanism reduces this number even further. Interestingly, experiments on the real robot produce even fewer schemas due to an additional constraint; the visual range of the cameras used on the real robot was more restricted than that in the simulator model.

Table 11 gives an example of a sequence of schemas learnt on the iCub, and the times of their creation⁸. Initially the robot has access to the primitive sensorimotor actions contained in the

learnt mappings, which include gazing and reaching. It also has preprogrammed reflex grasp and button-pressing actions.

At the outset the robot is presented with a green object. As the most novel stimuli this triggers available actions: first gaze, then reach. At 0:18, the robot receives tactile feedback from this action, which results in the generation of a new schema for touching a green object at that location. The excitation of this schema causes the action to be repeated. Noise in the system creates some subtle variation, and leads to the generation of some similar schemas. At 0:50 these are generalized into a schema for touching any colored object at any location (note that in this experiment we have set the requirements for generalization to a minimum to enable fast learning). The robot then tests the generalization by repeating the action. At 1:45 the excitation of the touching schemas have dwindled, and the grasp action now has the highest excitation. This is due to the similarity between the existing touch sensation and the recorded touch sensation triggered by closing the hand on itself. At 1:56 the robot generates a new schema for grasping a green object at that location, and this is quickly followed by the generalized version due to the similarity with the existing generalized touching schema. The robot cannot re-grasp, and so reaching again becomes the most excited action. At 2:19 the robot has moved its hand to a new position whilst still holding the object. This creates a new schema for moving an object that, following more repetition, becomes generalized at 2:36. At 3:32,

Table 10 | Effect of development and generalization on schema production.

Scenario	Number of schemas produced
Generalization only (Simulated Robot)	19,244
Stages only (Simulated Robot)	347
Stages, generalization (Simulated Robot)	227
Stages, generalization (Physical Robot)	115

⁸A video of this sequence can be seen at <http://youtu.be/3zb88qYmxMw>

Table 11 | Schema discovery on the iCub.

Time (mm:ss)	Preconditions	Action	Postconditions	Description
00:18	Green object at (17.5, 72.4)	Reach to (17.5, 72.4)	Hand at (17.5, 72.4) Green object at (17.5, 72.4) Touch sensation	New touch schema
00:50	\$z color object at (\$x,\$y)	Reach to (\$x,\$y)	Hand at (\$x,\$y) \$z color object at (\$x,\$y) Touch sensation	Generalized touch schema
01:56	Green object at (17.5, 72.4) Touch sensation	Grasp	Hand at (17.5, 72.4) Green object at (17.5, 72.4) Holding object	New grasping schema
02:01	\$z color object at (\$x,\$y) Touch sensation	Grasp	Hand at (\$x,\$y) \$z color object at (\$x,\$y) Holding object	Generalized grasp schema
02:19	Hand at (17.5, 72.4) Green object at (17.5, 72.4) Holding object	Reach to (8.8, 62.6)	Hand at (8.8, 62.6) Green object at (8.8, 62.6) Holding object	New transport schema
02:36	Hand at (\$x,\$y) Green object at (\$x,\$y) Holding object	Reach to (\$u,\$v)	Hand at (\$u,\$v) Green object at (\$u,\$v) Holding object	Generalized transport schema
03:42	Hand at (17.5, 72.4) Green object at (17.5, 72.4) Holding object	Release	Hand at (17.5, 72.4) Green object at (17.5, 72.4) Touch sensation	New release schema

The \$ notation specifies variable bindings, in left to right order.

after further repetition, the most excited option becomes the press action. This is particularly interesting as the robot is still holding the object, and provides the opportunity for learning basic tool use. However, in this instance the action does not cause a change in the world state, so no schema is generated. Finally, the release action becomes most exciting, and so the robot drops the object, learning the “release” schema.

4. DISCUSSION

This paper has described a longitudinal experiment in robotic developmental learning. Starting from a state of uncontrolled motor babbling, the iCub robot displayed a developmental progression, passing through several distinct behavioral stages, until skilled visio-motor behavior, involving reaching and manipulation of objects, was achieved. The result tables show the various learning times required for the robot to reach repeatable performance with reasonable accuracy and the total time for the whole process is less than 4 h. Such fast learning rates are crucial in real robot systems where online learning is essential, and training through many thousands of action cycles is quite impossible. This performance, which is typical of all our experiments, show that developmental learning algorithms offer serious potential for future real-time autonomous robots that must cope with novel events.

Whilst most comparable work in developmental robotics is focused on mechanisms within a single developmental stage, we are investigating longitudinal development and the transitions between multiple stages of behavior. Our methodology has been to implement the various subsystems in a way that facilitates their interaction, guided by the psychological literature to provide insight and inspiration. By closely following the stages evident in infancy, we find that learning in the robot is well directed along a trajectory that simplifies and reduces the amount of learning required. Furthermore, just as with infants, these trajectories will be similar but never exactly the same. In the early stages, where sensor and motor activity is being coordinated, learning is affected by variation in stimuli and motor babbling. In the later stages, the trajectory of schema development is dependent on the learnt primitive skills, the initial excitation of schemas, and the environment. Therefore trajectories can, and do, vary in their appearance across experiments.

Imposing carefully selected constraints can very effectively reduce the complexity of learning at each stage, with earlier stages providing valuable data for bootstrapping later stages. The experiments show some of the conditions for the interaction of different constraints (maturational or environmental) that can enhance learning rates. Whilst we have imposed the general order of constraints to structure the earlier stages of development, their release has been determined by internal measures, allowing the variations described above. Furthermore, we have shown how the early and late release of constraints impacts on development. We have also shown how stages may emerge based on environmental factors, or through experience, as shown by our schema experiments. Although it is possible to trigger constraint release by various means, as is sometimes necessary in experiment, we believe that emergent states based on current levels of development may

account for this process without recourse to specific mechanisms. This requires further investigation.

In following the longitudinal approach it becomes necessary to recognize that any current stage under study is conditioned by the previous stages which may feed in structures and experience that can influence the resultant performance. This means the earliest stage possible should be the start point and although we originally started with the newborn we realized that the fetal stage can make an important contribution in the bootstrapping sense. It seems the early sensorimotor organization occurring at this stage could be of considerable significance for the development of later abilities.

We have drawn on various sources for guidance on these issues, these include: the emergence of stereotypical movements and actions in the prenatal period (Mori and Kuniyoshi, 2010; Yamada and Kuniyoshi, 2012); learning to control saccades and gaze shifts (Srinivasa and Grossberg, 2008); and the emergence of stereotypical reaching behavior (Schlesinger et al., 2000), including the benefits of the ordered release of constraints (Savastano and Nolfi, 2012). Other relevant work includes that of Grupen and colleagues who were amongst the first to recognize the potential of the cephalocaudal progression of infant development as a robotic technique (Coelho et al., 2001; Grupen, 2005; Hart and Grupen, 2013), and the proximo-distal heuristic has been widely recognized, e.g., Elman (1993). Other key projects are investigating periods of cognitive growth through a variety of robotic platforms and models (Asada et al., 2009; Mori and Kuniyoshi, 2010), and reaching has received particular attention regarding the staged release of constraints (Savastano and Nolfi, 2012), their impact (Ramirez-Contla et al., 2012), and possible emergence (Stulp and Oudeyer, 2012).

Another distinct feature of the results is the use of motor babbling behavior to drive learning. Whilst most other similar systems use goal-driven and error-reducing methods, we note that goals and errors are usually specified by the system designers. We consider it important to investigate general action and open-ended exploratory/goal-finding behavior. In this context goals and errors are to be discovered or given significance by the agent itself. The simple novelty algorithm combined with motor babbling provides an effective exploratory learning mechanism that generates much pertinent data for learning about sensorimotor experience. Motor babbling is a form of spontaneous action and the excitation method applies to both single actions and action sequences during schema selection. This means that novel action patterns can emerge, as seen in the experiment, and this type of behavior is very reminiscent of infant play, which is also an exploratory goal-free behavior. Play has been long recognized as a critical and integral part of child development and the importance of novelty-driven play in infant development is well established (Bruner et al., 1976). We view play as an extension of motor babbling behavior, and schemas as the substrate to support this process. This hypothesis is described further in Lee (2011).

To summarise, this experiment has provided a demonstration of longitudinal development as a particularly fast and effective sensory-motor learning technique. Constraints have been used to shape infant-like behavior development, and we find these have an important role in speeding learning in robotic models. In

particular, data learnt at a more constrained stage can bootstrap later learning, leading to improved performance. Finally, the combination of very simple novelty detection mechanisms and intrinsic babbling algorithms are, at least, sufficient to drive learning of early sensory-motor coordination and basic skill acquisition.

ACKNOWLEDGMENTS

This research has received funds from the European Commission 7th Framework Programme (FP7/2007-2013), “Challenge 2—Cognitive Systems, Interaction, Robotics,” grant agreement No. ICT-IP-231722, project “IM-CLeVeR—Intrinsically Motivated Cumulative Learning Versatile Robots.”

REFERENCES

- Acredolo, L. P., and Evans, D. (1980). Developmental changes in the effects of landmarks on infant spatial behavior. *Dev. Psychol.* 16, 312. doi: 10.1037/0012-1649.16.4.312
- Asada, M., Hosoda, K., Kuniyoshi, Y., Ishiguro, H., Inui, T., Yoshikawa, Y., et al. (2009). Cognitive developmental robotics: a survey. *IEEE Trans. Auton. Ment. Dev.* 1, 12–34. doi: 10.1109/TAMD.2009.2021702
- Bayley, N. (1936). The development of motor abilities during the first three years: a study of sixty-one infants tested repeatedly. *Monogr. Soc. Res. Child Dev.* 1, 1–26. doi: 10.2307/1165480
- Berthier, N. E., and Carrico, R. L. (2010). Visual information and object size in infant reaching. *Infant Behav. Dev.* 33, 555–566. doi: 10.1016/j.infbeh.2010.07.007
- Berthier, N. E., Clifton, R. K., McCall, D. D., and Robin, D. J. (1999). Proximodistal structure of early reaching in human infants. *Exp. Brain Res.* 127, 259–269. doi: 10.1007/s002210050795
- Berthier, N. E., and Keen, R. (2006). Development of reaching in infancy. *Exp. Brain Res.* 169, 507–518. doi: 10.1007/s00221-005-0169-9
- Braitenberg, V., and Schüz, A. (1991). *Anatomy of the Cortex: Statistics and Geometry*. Berlin: Springer-Verlag. doi: 10.1007/978-3-662-02728-8
- Bremner, A. J., Holmes, N. P., and Spence, C. (2008). Infants lost in (peripersonal) space? *Trends Cogn. Sci.* 12, 298–305. doi: 10.1016/j.tics.2008.05.003
- Bremner, J. G. (1994). *Infancy*. Cambridge, MA: Blackwell.
- Bruner, J. (1990). *Acts of Meaning*. Cambridge, MA: Harvard University Press.
- Bruner, J. S. (1968). *Processes of Cognitive Growth: Infancy*. Worcester, MA: Clark University Press.
- Bruner, J. S., Jolly, A., and Sylva, K. (1976). *Play: Its Role in Development and Evolution*. New York: Basic Books.
- Butterworth, G., and Harris, M. (1994). *Principles of Developmental Psychology*. Hove: Lawrence Erlbaum Associates.
- Caligiore, D., Ferrautto, T., Parisi, D., Accornero, N., Capozza, M., and Baldassarre, G. (2008). “Using motor babbling and hebb rules for modeling the development of reaching with obstacles and grasping,” in *Proceedings of International Conference on Cognitive Systems COGSYS 2008*, eds R. Dillmann, C. Maloney, G. Sandini, T. Asfour, G. Cheng, and G. Metta (Karlsruhe: Springer), E1–E8.
- Cangelosi, A., Metta, G., Sagerer, G., Nolfi, S., Nehaniv, C., Fischer, K., et al. (2010). Integration of action and language knowledge: a roadmap for developmental robotics. *IEEE Trans. Auton. Ment. Dev.* 2, 167–195. doi: 10.1109/TAMD.2010.2053034
- Caporale, N., and Dan, Y. (2008). Spike timing-dependent plasticity: a hebbian learning rule. *Annu. Rev. Neurosci.* 31, 25–46. doi: 10.1146/annurev.neuro.31.060407.125639
- Carvalho, R., Tudella, E., and Savelsbergh, G. (2007). Spatio-temporal parameters in infant's reaching movements are influenced by body orientation. *Infant Behav. Dev.* 30, 26–35. doi: 10.1016/j.infbeh.2006.07.006
- Clifton, R. K., Muir, D. W., Ashmead, D. H., and Clarkson, M. G. (1993). Is visually guided reaching in early infancy a myth? *Child Dev.* 64, 1099–1110. doi: 10.2307/1131328
- Clifton, R. K., Rochat, P., Robin, D. J., and Bertheir, N. E. (1994). Multimodal perception in the control of infant reaching. *J. Exp. Psychol. Hum. Percept. Perform.* 20, 876–886. doi: 10.1037/0096-1523.20.4.876
- Coelho, J., Piater, J., and Grupen, R. (2001). Developing haptic and visual perceptual categories for reaching and grasping with a humanoid robot. *Robot. Auton. Syst.* 37, 195–218. doi: 10.1016/S0921-8890(01)00158-0
- De Vries, J. I. P., Visser, G. H. A., and Prechtl, H. F. R. (1984). Fetal motility in the first half of pregnancy. in *Continuity of Neural Function from Prenatal to Postnatal Life*, ed. H. F. R. Prechtl (London: Spastics International Medical Publications), 46–64.
- Earland, K., Law, J., Shaw, P., and Lee, M. (2014). Overlapping structures in sensory-motor mappings. *PLoS ONE* 9:e84240. doi: 10.1371/journal.pone.0084240
- Elman, J. L. (1993). Learning and development in neural networks: the importance of starting small. *Cognition* 48, 71–99. doi: 10.1016/0010-0277(93)90058-4
- Ennouri, K., and Bloch, H. (1996). Visual control of hand approach movements in new-borns. *Br. J. Dev. Psychol.* 14, 327–338. doi: 10.1111/j.2044-835X.1996.tb00709.x
- Fiorentino, M. R. (1981). *A Basis for Sensorimotor Development, Normal and Abnormal: The Influence of Primitive, Postural Reflexes on the Development and Distribution of Tone*. Springfield, IL: Charles C Thomas.
- Freedman, E. (2008). Coordination of the eyes and head during visual orienting. *Exp. Brain Res.* 190, 369–387. doi: 10.1007/s00221-008-1504-8
- Fuke, S., Ogino, M., and Asada, M. (2009). Acquisition of the head-centered peripersonal spatial representation found in vip neuron. *IEEE Trans. Auton. Mental Dev.* 1, 131–140. doi: 10.1109/TAMD.2009.2031013
- Gandhi, N. J., and Katmani, H. A. (2011). Motor functions of the superior colliculus. *Annu. Rev. Neurosci.* 34, 205–231. doi: 10.1146/annurev-neuro-061010-113728
- Girard, B., and Berthoz, A. (2005). From brainstem to cortex: computational models of saccade generation circuitry. *Prog. Neurobiol.* 77, 215–251. doi: 10.1016/j.pneurobio.2005.11.001
- Goodkin, F. (1980). The development of mature patterns of head –eye coordination in the human infant. *Early Hum. Dev.* 4, 373–386. doi: 10.1016/0378-3782(80)90042-0
- Goossens, H. H. L. M., and van Opstal, A. J. (1997). Human eye-head coordination in two dimensions under different sensorimotor conditions. *Exp. Brain Res.* 114, 542–560. doi: 10.1007/PL00005663
- Grupen, R. (2005). A developmental organization for robot behavior. DTIC Document ADA439115, Department of Computer Science, Massachusetts University, Amherst.
- Guerin, F., Kruger, N., and Kraft, D. (2013). A survey of the ontogeny of tool use: from sensorimotor experience to planning. *IEEE Trans. Auton. Mental Dev.* 5, 18–45. doi: 10.1109/TAMD.2012.2209879
- Guitton, D., and Volle, M. (1987). Gaze control in humans: eye-head coordination during orienting movements to targets within and beyond the oculomotor range. *J. Neurophysiol.* 58, 427–459.
- Hart, S., and Grupen, R. (2013). “Intrinsically motivated affordance discovery and modeling,” in *Intrinsically Motivated Learning in Natural and Artificial Systems*, eds G. Baldassarre and M. Mirolli (Berlin: Springer), 279–300. doi: 10.1007/978-3-642-32375-1_12
- Kaas, J. H. (1997). Topographic maps are fundamental to sensory processing. *Brain Res. Bull.* 44, 107–112. doi: 10.1016/S0361-9230(97)00094-4
- Kalnins, I., and Bruner, J. (1973). The coordination of visual observation and instrumental behavior in early infancy. *Perception* 2, 307–314. doi: 10.1080/p020307
- Kaufman, J., Gilmore, R. O., and Johnson, M. H. (2006). Frames of reference for anticipatory action in 4-month-old infants. *Infant Behav. Dev.* 29, 322–333. doi: 10.1016/j.infbeh.2005.01.003
- Kuniyoshi, Y., and Sangawa, S. (2006). Early motor development from partially ordered neural-body dynamics: experiments with a cortico-spinal-musculo-skeletal model. *Biol. Cybern.* 95, 589–605. doi: 10.1007/s00422-006-0127-z
- Law, J., Lee, M., Hülse, M., and Tomassetti, A. (2011). The infant development timeline and its application to robot shaping. *Adapt. Behav.* 19, 335–358. doi: 10.1177/1059712311419380
- Law, J., Shaw, P., and Lee, M. (2013). A biologically constrained architecture for developmental learning of eye-head gaze control on a humanoid robot. *Auton. Robot.* 35, 77–92. doi: 10.1007/s10514-013-9335-2
- Law, J., Shaw, P., Lee, M., and Sheldon, M. (2014). “From saccades to play: a model of coordinated reaching through simulated development on a humanoid robot,” in *IEEE Transactions on Autonomous Mental Development*.
- Lee, M., Meng, Q., and Chao, F. (2007). Staged competence learning in developmental robotics. *Adapt. Behav.* 15, 241–255. doi: 10.1177/1059712307082085

- Lee, M. H. (2011). "Intrinsic activity: from motor babbling to play," in *2011 IEEE International Conference on Development and Learning (ICDL)*. Vol. 2 (Frankfurt am Main), 1–6. doi: 10.1109/DEVLRN.2011.6037375
- Lungarella, M., Metta, G., Pfeifer, R., and Sandini, G. (2003). Developmental robotics: a survey. *Connect. Sci.* 15, 151–190. doi: 10.1080/09540090310001655110
- Mallot, H., Von Seelen, W., and Giannakopoulos, F. (1990). Neural mapping and space-variant image processing. *Neural Netw.* 3, 245–263. doi: 10.1016/0893-6080(90)90069-W
- Markram, H., Gerstner, W., and Sjöström, P. J. (2011). A history of spike-timing-dependent plasticity. *Front. Synaptic Neurosci.* 3:4. doi: 10.3389/fnsyn.2011.00004
- Maurer, D., and Maurer, C. (1988). *The World of the Newborn*. New York: Basic Books.
- Morasso, P., and Sanguineti, V. (1995). Self-organizing body schema for motor planning. *J. Motor Behav.* 27, 52–66. doi: 10.1080/00222895.1995.9941699
- Mori, H., and Kuniyoshi, Y. (2010). "A human fetus development simulation: Self-organization of behaviors through tactile sensation," in *2010 IEEE 9th International Conference on Development and Learning* (Ann Arbor, MI), 82–87. doi: 10.1109/DEVLRN.2010.5578860
- Nagai, Y., Asada, M., and Hosoda, K. (2006). Learning for joint attention helped by functional development. *Adv. Robot.* 20, 1165–1181. doi: 10.1163/156855306778522497
- Nagai, Y., Kawai, Y., and Asada, M. (2011). "Emergence of mirror neuron system: Immature vision leads to self-other correspondence," in *2011 IEEE International Conference on Development and Learning*. Vol. 2 (Frankfurt am Main), 1–6. doi: 10.1109/DEVLRN.2011.6037335
- Natale, L., Nori, F., Metta, G., Fumagalli, M., Ivaldi, S., Pattacini, U., et al. (2013). The icub platform: a tool for studying intrinsically motivated learning. in *Intrinsically Motivated Learning in Natural and Artificial Systems*, eds G. Baldassarre and M. Mirolli (Berlin: Springer), 433–458. doi: 10.1007/978-3-642-32375-1_17
- Needham, A., Barrett, T., and Peterman, K. (2002). A pick-me-up for infants' exploratory skills: Early simulated experiences reaching for objects using 'sticky mittens' enhances young infants' object exploration skills. *Infant Behav. Dev.* 25, 279–295. doi: 10.1016/S0163-6383(02)00097-8
- Newcombe, N. S., and Huttenlocher, J. (2006). "Development of spatial cognition," in *Handbook of Child Psychology*, eds D. Kuhn, R. S. Siegler, W. Damon, and R. M. Lerner (Hoboken, NJ: Wiley Online Library). doi: 10.1002/9780470147658.chpsy0217
- Piaget, J. (1973). *The Child's Conception of the World*. London: Paladin.
- Piaget, J., and Cook, M. (1952). *The Origins of Intelligence in Children*. New York, NY: WW Norton & Co. doi: 10.1037/11494-000
- Piaget, J., and Inhelder, B. (1956). *The Child's Conception of Space*. London: Routledge and Kegan Paul.
- Ramirez-Contla, S., Cangelosi, A., and Marocco, D. (2012). "Developing motor skills for reaching by progressively unlocking degrees of freedom on the icub humanoid robot," in *Proceedings of the Post-Graduate Conference on Robotics and Development of Cognition* (Lausanne). doi: 10.2390/biec02-robotdoc2012-14
- Redgrave, P., Gurney, K., Stafford, T., Thirkettle, M., and Lewis, J. (2013). "The role of the basal ganglia in discovering novel actions," in *Intrinsically Motivated Learning in Natural and Artificial Systems*, eds G. Baldassarre and M. Mirolli (Berlin: Springer), 129–150. doi: 10.1007/978-3-642-32375-1_6
- Rochat, P. (1993). "Hand-mouth coordination in the newborn: morphology, determinants, and early development of a basic act," in *The Development of Coordination in Infancy. Advances in Psychology*, Vol. 97, ed G. J. P. Savelsbergh (Amsterdam: North-Holland), 265–288.
- Saegusa, R., Metta, G., Sandini, G., and Sakka, S. (2009). Active motor babbling for sensorimotor learning. in *IEEE International Conference on Robotics and Biomimetics, 2008* (Bangkok), 794–799. doi: 10.1109/ROBIO.2009.4913101
- Savastano, P., and Nolfi, S. (2012). Incremental learning in a 14 dof simulated icub robot: modeling infant reach/grasp development. *Biomimet. Biohybrid Syst.* 7375, 250–261. doi: 10.1007/978-3-642-31525-1_22
- Schlesinger, M., Parisi, D., and Langer, J. (2000). Learning to reach by constraining the movement search space. *Dev. Sci.* 3, 67–80. doi: 10.1111/1467-7687.00101
- Shaw, P., Law, J., and Lee, M. (2012). An evaluation of environmental constraints for biologically constrained development of gaze control on an icub robot. *Paladyn* 3, 147–155. doi: 10.2478/s13230-013-0103-y
- Shaw, P., Law, J., and Lee, M. (2014). A comparison of learning strategies for biologically constrained development of gaze control on an icub robot. *Auton. Robot.* doi: 10.1007/s10514-013-9378-4
- Sheldon, M. T. (2012). *Intrinsically Motivated Developmental Learning of Communication in Robotic Agents*. Ph.D. thesis, Aberystwyth University.
- Sheldon, M. T., and Lee, M. (2011). "PSchema: a developmental schema learning framework for embodied agents," in *Proceedings of the IEEE Joint International Conference on Development and Learning, and Epigenetic Robotics* (Frankfurt).
- Sheridan, M. D. (1973). *From Birth to Five Years*. (Windsor: NFER Publishing). doi: 10.4324/9780203273586
- Shirley, M. M. (1933). *The First Two Years - A Study of Twenty-Five Babies. Intellectual Development*, Vol. 2 (Minneapolis, MN: University of Minnesota Press).
- Srinivas, N., and Grossberg, S. (2008). A head-neck-eye system that learns fault-tolerant saccades to 3-d targets using a self-organizing neural model. *Neural Netw.* 21, 1380–1391. doi: 10.1016/j.neunet.2008.07.007
- Stoytchev, A. (2009). Some basic principles of developmental robotics. *IEEE Trans. Auton. Mental Dev.* 1, 122–130. doi: 10.1109/TAMD.2009.2029989
- Stulp, F., and Oudeyer, P.-Y. (2012). "Emergent proximo-distal maturation through adaptive exploration," in *IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL), 2012*, 1–6. doi: 10.1109/DevLrn.2012.6400586
- Vasilyeva, M., and Lourenco, S. F. (2012). Development of spatial cognition. *Wiley Interdiscip. Rev. Cogn. Sci.* 3, 349–362. doi: 10.1002/wcs.1171
- Vernon, D., Hofsten, C., and Fadiga, L. (2010). *A Roadmap for Cognitive Development in Humanoid Robots*. Cognitive Systems Monographs (COSMOS). Vol. 11 (Berlin: Springer).
- von Hofsten, C., and Rönnqvist, L. (1993). The structuring of neonatal arm movements. *Child Dev.* 64, 1046–1057. doi: 10.2307/1131326
- White, B. L., Castle, P., and Held, R. (1964). Observations on the development of visually-directed reaching. *Child Dev.* 35, 349–364. doi: 10.1111/j.1467-8624.1964.tb05944.x
- Yamada, Y., and Kuniyoshi, Y. (2012). Embodiment guides motor and spinal circuit development in vertebrate embryo and fetus. In *IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL)*, 1–6. doi: 10.1109/DevLrn.2012.6400578

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 05 July 2013; **accepted:** 03 January 2014; **published online:** 27 January 2014.
Citation: Law J, Shaw P, Earland K, Sheldon M and Lee M (2014) A psychology based approach for longitudinal development in cognitive robotics. *Front. Neurorobot.* 8:1. doi: 10.3389/fnbot.2014.00001

This article was submitted to the journal *Frontiers in Neurorobotics*.

Copyright © 2014 Law, Shaw, Earland, Sheldon and Lee. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



A game theoretic framework for incentive-based models of intrinsic motivation in artificial systems

Kathryn E. Merrick * and Kamran Shafi

School of Engineering and Information Technology, University of New South Wales, Australian Defence Force Academy, Canberra, ACT, Australia

Edited by:

Tom Stafford, University of Sheffield, UK

Reviewed by:

Christian C. Luhmann, Stony Brook University, USA

Chrisantha T. Fernando, University of Sussex, UK

***Correspondence:**

Kathryn E. Merrick, School of Engineering and Information Technology, University of New South Wales, Australian Defence Force Academy, Northcott Drive, Canberra, ACT 2600, Australia
e-mail: k.merrick@adfa.edu.au

An emerging body of research is focusing on understanding and building artificial systems that can achieve open-ended development influenced by intrinsic motivations. In particular, research in robotics and machine learning is yielding systems and algorithms with increasing capacity for self-directed learning and autonomy. Traditional software architectures and algorithms are being augmented with intrinsic motivations to drive cumulative acquisition of knowledge and skills. Intrinsic motivations have recently been considered in reinforcement learning, active learning and supervised learning settings among others. This paper considers game theory as a novel setting for intrinsic motivation. A game theoretic framework for intrinsic motivation is formulated by introducing the concept of optimally motivating incentive as a lens through which players perceive a game. Transformations of four well-known mixed-motive games are presented to demonstrate the perceived games when players' optimally motivating incentive falls in three cases corresponding to strong power, affiliation and achievement motivation. We use agent-based simulations to demonstrate that players with different optimally motivating incentive act differently as a result of their altered perception of the game. We discuss the implications of these results both for modeling human behavior and for designing artificial agents or robots.

Keywords: intrinsic motivation, game theory, agents, prisoner's dilemma, leader, chicken, battle of the sexes

INTRODUCTION

Game theory is the study of strategic decision-making (Guillermo, 1995). It has been used to study a variety of human and animal behaviors in economics, political science, psychology, biology, and other areas. Game theoretic approaches have also been utilized in robotics for tasks such as multi-robot coordination and optimization (Meng, 2008; Kaminka et al., 2010) as well as for analyzing and implementing behavior in software agents (Parsons and Wooldridge, 2002). This paper presents a game theoretic framework for intrinsic motivation and considers how motivation might drive cultural learning during strategic interactions. The work provides stepping stones toward intrinsically motivated, game theoretic approaches to modeling strategic interactions. Potential applications include the study of human behavior or modeling open-ended development in robots or artificial agents.

In humans, individual differences in the strength of motives such as power, achievement and affiliation have been shown to have a significant impact on behavior in social dilemma games (Terhune, 1968; Kuhlman and Marshello, 1975; Kuhlman and Wimberley, 1976; Van Run and Liebrand, 1985) and during other kinds of strategic interactions (Atkinson and Litwin, 1960). Some models of these phenomena exist for artificial agents (Simkins et al., 2010; Merrick and Shafi, 2011), but these models have not yet been widely studied for strategic interactions, competition and cooperation between artificial agents.

This paper presents a game theoretic approach to modeling differences in decision-making between individuals caused

by differences in their perception of the payoff during certain strategic interactions. Specifically we consider cases where differences in perception are caused by different motivational preferences held by individuals. We study strategic decision-making in the context of mixed-motive games. Four archetypical two-by-two mixed-motive games are considered: prisoner's dilemma (PD), leader, chicken, and battle-of-the-sexes (BoS) (Rapoport, 1967; Colman, 1982). We introduce the concept of optimally motivating incentive and demonstrate that agents with different optimally motivating incentives perceive the four games differently. We show that the perceived games have different Nash Equilibrium (NE) points (Nash, 1950) to the original games. This causes agents with different optimally motivating incentives to act differently. We discuss the implications of these results both for modeling human behavior and for designing artificial agents or robots with certain behavioral characteristics.

In the remainder of this Section, section Mixed-Motive Games introduces mixed-motive games and section Solution Strategies for Mixed-Motive Games reviews relevant existing models of strategic decision-making. Section Solution Strategies for Mixed-Motive Games also discusses the specific contributions of this paper in that context and introduces the background formal notations used in the rest of the paper. Section Incentive-Based Models of Motivation reviews literature from motivational psychology about the influence of incentive-based motivation on decision-making as inspiration for the new models in sections Materials and Methods. Sections Materials and Methods introduces our new notation for incentives and shows how each of

the four mixed-motive games are transformed into various new games when different optimally motivating incentives are chosen for agent players. Section Results presents a suite of agent-based simulations demonstrating that players with different optimally motivating incentive act differently as a result of their altered perception of the game. We conclude in section Discussion with a discussion of the implications of the work and future directions it may take.

MIXED-MOTIVE GAMES

This paper will consider two-player mixed motive games with the generic structure shown in Matrix 1. Each player, (Player 1 and Player 2) has a choice of two actions: *C* or *D*. Depending on the combination of actions chosen by both players, Player 1 is assigned a payoff value V_1 and Player 2 is assigned a payoff value V_2 . V_1 and V_2 can have values of *T*, *R*, *P*, or *S*. The value *R* is the reward if both players choose *C*. In other words, *R* is the reward for a (*C*, *C*) outcome. *P* is the punishment if both players defect [joint *D* choices leading to a (*D*, *D*) outcome]. In a mixed-motive game, *P* must be less than *R*. *T* represents the temptation to defect (choose action *D*) from the (*C*, *C*) outcome and thus, in a mixed-motive game *T* must be greater than *R*. Finally, *S* is the sucker's payoff for choosing *C* when the other player chooses *D*.

Formally, the game **G** presents players with a payoff matrix:

$$\mathbf{G} = \begin{bmatrix} P & T \\ S & R \end{bmatrix}$$

The generic game **G** can be used to define a number of specific games by fixing the relationships between *T*, *R*, *P*, and *S*. Four well-known two-by-two mixed motive games and the relationships that define them are (Colman, 1982):

1. Prisoner's Dilemma: $T > R > P > S$
2. Leader: $T > S > R > P$
3. Chicken: $T > R > S > P$
4. Battle of the Sexes: $S > T > R > P$

A number of variations of these games do exist (as well as other distinct games), but this paper will focus on the four games as defined above.

Matrix 1. A generic two-by-two mixed-motive game **G**. *T* must be greater than *R* and *R* must be greater than *P*.

		Player 2	
Player 1		<i>D</i>	<i>C</i>
	<i>D</i>	<i>P, P</i>	<i>T, S</i>
	<i>C</i>	<i>S, T</i>	<i>R, R</i>

The PD game (Rapoport and Chammah, 1965; Poundstone, 1992) is perhaps the most well-known of the four games studied in this paper. It derives its name from a hypothetical strategic interaction in which two people are arrested for involvement in a crime. They are held in separate cells and cannot communicate with each other. The police have insufficient evidence for a conviction unless at least one of the prisoners discloses certain incriminating information. Each prisoner has a choice between concealing information from the police (action *C*) or disclosing it

(action *D*). If both conceal, both will be acquitted and the payoff to both will be $V_1 = V_2 = R$. If both disclose, both will be convicted and receive minor punishments: $V_1 = V_2 = P$. If only one prisoner discloses information he will be acquitted and, in addition, receive a reward for his information. In this case, the prisoner who conceals information will receive a heavy punishment. For example if Player 1 discloses and Player 2 conceals, the payoffs will be $V_1 = T$ and $V_2 = S$. Player 2 in this situation is sometimes referred to as the "martyr" because he generates the highest payoff for the other player and the lowest payoff for himself.

The PD game has been used as a model for arms races, voluntary wage restraint, conservation of scarce resources and the iconic "tragedy of the commons" (see Colman, 1982, for a review). More recently, however, biologists have argued that individual variation in motivation and perception means that a majority of strategic interactions do not, in fact, conform to the PD model (Johnson et al., 2002). The models presented in our paper demonstrate one possible explanation for this latter view. Specifically, they show how a valid PD matrix can be transformed into another game that no longer represents a PD scenario as a result of individuals having different motives.

The game of Leader (Rapoport, 1967) is an analogy for real-world interactions such as those between pedestrians or drivers in traffic. For example, suppose two pedestrians wish to enter a turnstile. Each must decide whether to walk into the turnstile first (action *D*) or concede right of way and wait for the other to walk in (action *C*). If both pedestrians wait, then both will be delayed and receive payoffs $V_1 = V_2 = R$. If they both decide to walk first, a socially awkward situation results in the worst payoff $V_1 = V_2 = P$ to both. If one decides to walk and the other waits, the "leader" will be able to walk through unimpeded, receiving the highest payoff *T*, while the "follower" will be able to walk through afterwards giving the second best payoff *S*. Other examples of real world interactions abstracted by the Leader game include two drivers at opposite ends of a narrow, one-lane bridge, or two drivers about to merge from two lanes into one. In some such real-world situations there are rules of thumb that prevent the leader game from emerging, for example flashing headlights at a bridge to concede right of way. However, when such communication fails or is impossible, individuals' motivations have an influential role in decision-making and in how individuals interpret the scenario. We make the standard assumption that there is no communication between agents.

In the game of Chicken two motorists speed toward each other on a collision course. Each has the option of swerving to avoid a collision, and thereby showing themselves to be "chicken" (action *C*) or of driving straight ahead (action *D*). If both players are "chicken," each gets a payoff of $V_1 = V_2 = R$. If only one player is "chicken" and the other drives straight on, then the "chicken" loses face and the other player, the "exploiter," wins a prestige victory. For example if Player 1 is "chicken" and Player 2 drives, the payoffs will be $V_1 = S$ and $V_2 = T$. If both players drive a collision will occur and both players will receive the worst payoff $V_1 = V_2 = P$. The game of Chicken has also been used to model real-world scenarios in national and international politics involving bilateral threats, as well as animal conflicts and

Darwinian selection of evolutionarily stable strategies (Maynard-Smith, 1982).

Finally, the BoS game can be thought of as modeling a predicament between two friends with different interests in entertainment. Each prefers a certain form of entertainment that is different to the other, but both would rather go out together than alone. If both opt for their preferred entertainment, leading to a (C, C) outcome, then each ends up going alone and receiving a payoff of $V_1 = V_2 = R$. A worse outcome (D, D) results if both make the sacrifice of going to the entertainments they dislike as they both end up alone and $V_1 = V_2 = P$. If, however, one chooses their preferred entertainment and the other plays the role of “hero” and makes the sacrifice of attending the entertainment they dislike then the outcome is better for both of them (either $V_1 = T$ and $V_2 = S$ or $V_1 = S$ and $V_2 = T$). The payoff matrix for BoS is relatively similar to that of Leader, with the only difference in the definition being the relationship between T and S . In Leader $T > S$, while in BoS $S > T$. This reflects the real-world relationship that is often perceived between leadership and sacrifice (Van Knippenberg and Van Knippenberg, 2005). We will see in section Results that some of the game transformations that are perceived by agents using our model of optimally motivating incentive also reflect this relationship.

SOLUTION STRATEGIES FOR MIXED-MOTIVE GAMES

A strategy σ is a function that takes a game as input and outputs an action to perform according to some plan of play. This paper will focus on pure strategies, such as “always choose action C ” and mixed strategies that make a stochastic choice between two pure strategies with a fixed frequency. Suppose we denote the probability that Player 2 will choose action C as $P_2(C)$, then the expected payoff for the two pure strategies available to Player 1 (“always play C ” or “always play D ”) can be computed as follows:

$$E_1(C) = P_2(C)R + [1 - P_2(C)]S$$

$$E_1(D) = P_2(C)T + [1 - P_2(C)]P$$

Using this information, a player can choose the strategy with the maximum expected payoff. A variation on this idea that takes into account individual differences in preference is utility theory (Keeney and Raiffa, 1976; Glimcher, 2011). Utility theory acknowledges that the values of different outcomes for different people are not necessarily equivalent to their raw payoff values V . Formally, a utility function $U(V)$ is a twice differentiable function defined for $V > 0$ which has the properties of non-satiation [the first derivative $U'(V) > 0$] and risk aversion [the second derivative $U''(V) < 0$]. The non-satiation property implies that the utility function is monotonic, while the risk aversion property implies that it is concave. Utility theories were first proposed in the 1700s and have been developed and critiqued in a range of fields including economics (Kahneman and Tversky, 1979) and game theory (Von Neumann and Morgenstern, 1953).

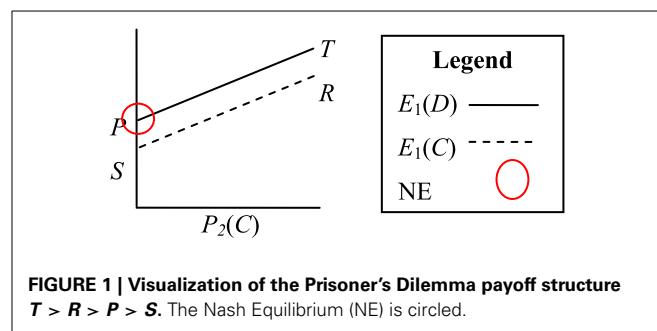
Alternatives have also been proposed to model effects that are inconsistent to utility theory. Examples include prospect theory (Kahneman and Tversky, 1979) and lexicographic preferences

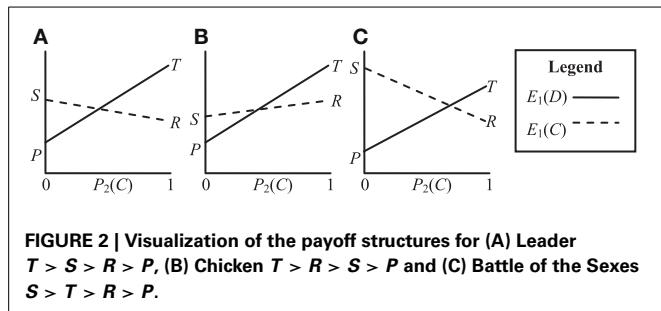
(Fishburn, 1974). The models in this paper can also be thought of as an alternative to utility theory that uses theories of motivation to determine how to compute individuals’ preferences. Various other techniques have been proposed to model decision-making under uncertainty, that is, when it is not possible to assign meaningful probabilities to alternative outcomes. Many of these techniques capture “rules of thumb” or heuristics used in human decision-making (Gigerenzer and Todd, 1999). Examples include the maximax, maximin, and regret principles.

The strategies chosen by players and their corresponding payoffs constitute a NE (Nash, 1950) if no player can benefit by changing their strategy while the other player keeps theirs unchanged. This latter definition covers mixed strategies M in which players make probabilistic random choices between actions. Formally, if we consider a pair of strategies, σ_1 and σ_2 , and denote the expected payoff for Player 1 using σ_1 against Player 2 using σ_2 as $E_1(\sigma_1, \sigma_2)$, then the two strategies are in equilibrium if $E_1(\sigma_1, \sigma_2) \geq E_1(\sigma'_1, \sigma_2)$ for all $\sigma'_1 \neq \sigma_1$. In other words, the strategies are in equilibrium if there is no alternative strategy for Player 1 that would improve Player 1’s expected payoff against Player 2 if Player 2 continues to use strategy σ_2 (Guillermo, 1995).

Suppose we consider the principles discussed above with reference to the four games described in section Mixed-Motive Games. In the PD game there is a pure strategy equilibrium point (D, D) from which neither player benefits from unilateral deviation, although both benefit from joint deviation. We can visualize this game in terms of expected payoff as shown in Figure 1. We denote the probability of Player 2 choosing C as $P_2(C)$, the expected payoff if Player 1 chooses D as $E_1(D)$, and the expected payoff for Player 1 choosing C as $E_1(C)$. The visualization shows that the definition of PD ($T > R > P > S$) implies that $E_1(D) > E_1(C)$ regardless of $P_2(C)$. In other words, the strategy of choosing D dominates the strategy of choosing C . The NE for this game (D, D) is shown circled in Figure 1.

In contrast to the PD game, the Leader, Chicken and BoS games all have $E_1(D) > E_1(C)$ for $P_2(C) = 1$ and $E_1(D) < E_1(C)$ for $P_2(C) = 0$. In other words, these games have two asymmetric equilibrium points (C, D) and (D, C) . However, neither of these equilibrium points is strongly stable because the players disagree about which is preferable. The three games do, however, have a mixed-strategy NE, meaning that players will tend to evolve strategies that choose C with some fixed probability. We can also visualize these games in terms of their expected payoff as shown





in **Figure 2**. The NE probability of players choosing C is defined by the point at which $E_1(D)$ and $E_1(C)$ intersect, i.e.:

$$E_1(C) = E_1(D)$$

$$[R - S]P_2(C) + S = [T - P]P_2(C) + P$$

$$P_2(C) = \frac{P - S}{R - S - T + P}$$

and likewise for $P_1(C)$.

Evolutionary game theory (Maynard-Smith, 1982) combines classical game theory with learning. Evolutionary dynamics predict the equilibrium outcomes of a multi-agent system when the individual agents use learning algorithms to choose actions in iterative game-play. Two-population replicator dynamics, for example, model learning when players may have different strategies. In this model, suppose we combine the probabilities of Player 1 playing C and D in a vector form $p = [p_C, p_D]$ such that $p_C = P_1(C)$ and $p_D = P_1(D)$ and the probabilities of Player 2 playing C and D $q = [q_C, q_D]$ such that $q_C = P_2(C)$ and $q_D = P_2(D)$. The replicator dynamics in this case are:

$$\Delta p_i = p_i[(\mathbf{G}q)_i - \mathbf{pG}^T q] \quad (1)$$

$$\Delta q_i = q_i[(\mathbf{pG}^T)_i - \mathbf{pG}^T \mathbf{q}^T] \quad (2)$$

where \mathbf{G} is the payoff matrix defined by the game being played. In this model, pure strategies tend to dominate over time and mixed-strategies are unstable.

In this paper, we use two-population replicator dynamics to model cultural learning (as opposed to biological evolution) when mixed-motive games are played iteratively. Borgers and Sarin (1997) showed that Cross' learning model for two players iteratively playing "habit forming games" converges to asymmetric continuous time replicator dynamics. Our approach is a stepping-stone toward simulating and analyzing strategic interactions between agents modeling known motive profiles.

While classical game theory discussed above offers a wide range of insights into behavior in strategic interactions, it is not necessarily designed to model human decision-making. In fact, there is evidence of humans not conforming to NE strategies in many kinds of strategic interaction (Terhune, 1968; McKelvey and Palfrey, 1992; Li et al., 2010). As a result, researchers have started to develop alternative approaches. The field of behavioral game theory (Camerer, 2003, 2004) is concerned with developing models of behavior under assumptions of bounded rationality. These

models take into account factors such as the heterogeneity of a population, the ability of individuals to learn and adapt during strategic interactions and the role of emotional and psychological factors in strategic decision-making. The purposes of this work fall into two broad categories: (1) to produce computational models that can explain and predict human behavior during strategic interactions that does not conform to classical game theoretic models (Valluri, 2006) and (2) to build artificial systems that can exhibit certain desirable behavioral characteristics such as cooperation or competitiveness (Sandholm and Crites, 1996; Claus and Boutilier, 1998; Vassiliades and Christodoulou, 2010), cooperation during strategic interactions (Valluri, 2006) and improved performance against human adversaries who also have bounded rationality and limited observation (Pita et al., 2010). The work in our paper differs from previous work in this area by its focus on the role of motivation in decision-making.

INCENTIVE-BASED MODELS OF MOTIVATION

In motivational psychology, incentive is defined as a situational characteristic associated with possible satisfaction of a motive (Heckhausen and Heckhausen, 2008). A range of incentive-based motivation theories exist, dealing with both internal and external incentives. Examples of internal incentives include the novelty, difficulty or complexity of a situation. Examples of external incentives include money and points or "payoff" in a game. For the remainder of this paper we define incentive I as a value that is proportional to payoff V defined in section Mixed-Motive Games. The key aspect of incentive-based motivation to be embedded in the game theoretic framework in this paper is that different individuals have different intrinsic preferences for incentives. These different intrinsic motivations cause individuals to perceive the payoff matrix specified by a game differently and act according to their own transformation of that matrix.

The following sub-sections describe three incentive-based models of motivation and the different motivational preferences they inspire. While we do not explicitly embed these models in our proposed game theoretic framework, they inform the cases of optimally motivating incentive and corresponding game transformations that we study in section Materials and Methods. The three motives considered are the "influential trio" proposed by Heckhausen and Heckhausen (2008): achievement, affiliation, and power motivation. These theories are the basis of competence-seeking behavior, relationship-building and resource-controlling behavior in humans.

Achievement motivation

Achievement motivation drives humans to strive for excellence by improving on personal and societal standards of performance. Perhaps the foremost psychological model of achievement motivation is Atkinson's Risk-Taking Model (RTM) (Atkinson, 1957). It defines achievement motivation in terms of conflicting desires to approach success or avoid failure. Six variables are used: incentive for success (equated with value of success); probability of success (equated with difficulty); strength of motivation to approach success; incentive for avoiding failure; probability of failure; and strength of motivation to avoid failure. Success motivated individuals perceive an inverse linear relationship between

incentive and probability of success (Atkinson and Litwin, 1960; Atkinson and Raynor, 1974). They tend to favor goals or actions with moderate incentives which can be interpreted as indicating a moderate probability of success or moderate difficulty. We examine the case of success-motivated individuals in this paper, by examining the case where individuals with a moderate optimally motivating incentive engage in strategic interactions.

Affiliation motivation

Affiliation refers to a class of social interactions that seek contact with formerly unknown or little known individuals and maintain contact with those individuals in a manner that both parties experience as satisfying, stimulating and enriching (Heckhausen and Heckhausen, 2008). The need for affiliation is activated when an individual comes into contact with another unknown or little known individual. While theories of affiliation have not been developed mathematically to the extent of the RTM, affiliation can be considered from the perspective of incentive and probability of success (Heckhausen and Heckhausen, 2008). In contrast to success-motivated individuals, individuals high in affiliation motivation may select goals with a higher probability of success and/or lower incentive. This often counter-intuitive preference can be understood as avoiding public competition and conflict. Affiliation motivation is thus an important balance to power motivation, but can also lead to individuals with high affiliation motivation underperforming their achievement motivated colleagues.

Power motivation

Power can be described as a domain-specific relationship between two individuals, characterized by the asymmetric distribution of social competence, access to resources or social status (Heckhausen and Heckhausen, 2008). Power is manifested by unilateral behavioral control and can occur in a number of different ways. Types of power include reward power, coercive power, legitimate power, referent power, expert power, and informational power. As with affiliation, power motivation can be considered with respect to incentive and probability of success. Specifically, there is evidence to indicate that the strength of satisfaction of the power motive depends solely on incentive and is unaffected by the probability of success (McClelland and Watson, 1973). Power motivated individuals select high-incentive goals, as achieving these goals gives them significant control of the resources and reinforcers of others.

Computational models of achievement, affiliation, and power motivation

Previous work has modeled incentive-based motivation functions computationally for agents with power, achievement, and affiliation motive profiles making one-off decisions (Merrick and Shafi, 2011). For example, **Figure 3** shows a possible computational motive profile as a sum of three curves for achievement, affiliation, and power motivation. Unlike utility functions, motivation functions may be non-monotonic and non-concave. The highest peak indicates the level of incentive I that produces the strongest resultant motivational tendency $m(I)$ for action. Assuming a

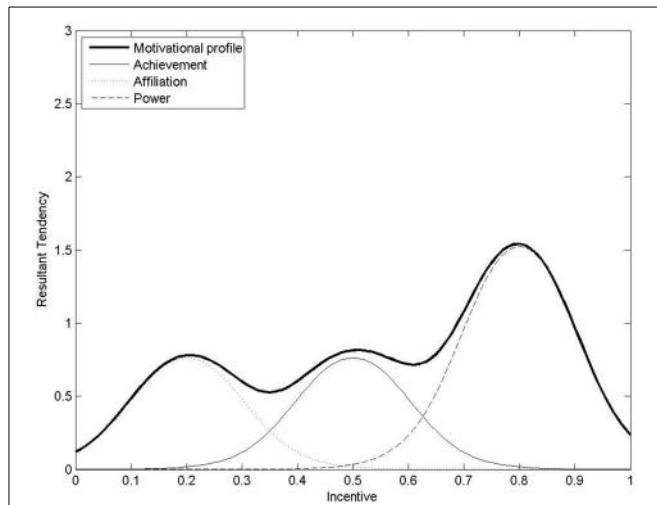


FIGURE 3 | A computational motive-profile as the sum of achievement, affiliation and power motivation. The resultant tendency for action is highest for incentive of 0.8 (the optimally motivating incentive for this agent). This agent may be qualitatively classified as “power-motivated” as its optimally motivating incentive is relatively high on the [0, 1] scale for incentive. Image from (Merrick and Shafi, 2011).

[0, 1] scale for incentive, agents are qualitatively classified as power, achievement or affiliation motivated if their optimally motivating incentive is high, moderate or low, respectively.

MATERIALS AND METHODS

The previous section establishes that individuals can view incentives differently. Broadly speaking, individuals with strong power, achievement, or affiliation may favor high, moderate, and low incentives, respectively. In a game theoretic setting this suggests that individuals may not play an explicitly described game, but rather act in response to their own idiosyncratic payoff matrix. This phenomenon is not captured by classical game theory or utility based models because of the non-monotonic and non-concave nature of motivation functions.

Our approach in this paper brings the idea of a non-monotonic intrinsic motivation function to game theory by modeling players as having different “optimally motivating incentives.” Optimally motivating incentives are scalar values that represent different motive profiles in a compressed form. Formally, suppose we have two agents A_1 and A_2 playing a mixed-motive game \mathbf{G} . We denote the optimally motivating incentive of A_1 as I_1^* and the optimally motivating incentive of A_2 as I_2^* . I_j^* is thus the value that maximizes the motivation function $m_j(I)$ of agent A_j . This paper is not concerned further with the definition of the function m . We focus instead on the game transformations that result from introducing I_j^* .

As we have seen, in a two-by-two game, there are four possible outcomes: (C, C) , (D, D) , (C, D) , and (D, C) . The incentive values for each possible outcome from the perspective of the player playing the first listed action are $I = R$, $I = P$, $I = S$, or $I = T$. (See section Mixed-Motive Games and Matrix 1.) Suppose each agent A_j wishes to adopt a strategy that results in an outcome that minimizes the difference between I and their individual optimally

motivating incentive I_j^* . That is, each agent wishes to minimize $|I - I_j^*|$. This means that agents with different values of I_j^* will perceive the incentives T, S, R , and P differently.

We define perceived incentive I'_j as a measure of the perceived value of a particular incentive I , for a particular agent A_j . If we further suppose that the maximum perceived incentive must be equal to the maximum incentive I_{\max} in the original game, then we can formalize the notion of perceived incentive I'_j as:

$$I'_j = I_{\max} - |I - I_j^*|$$

That is, perceived incentive is equal to maximum incentive minus the error between actual and optimal incentive. This means that I_{\max} only has the highest perceived value if it is closest to the agent's optimally motivating incentive I_j^* . In practice the implications are that each incentive I will be perceived differently by agents with different optimally motivating incentives I_j^* . In addition, the highest actual incentive may not be the highest perceived incentive for all agents.

We can now define the perceived incentives T', P', S' , and R' of each incentive in the original game. In PD, Leader, and Chicken the maximum incentive is $I_{\max} = T$ so we have:

$$\begin{aligned} T'_j &= T - |T - I_j^*| & R'_j &= T - |R - I_j^*| \\ P'_j &= T - |P - I_j^*| & S'_j &= T - |S - I_j^*| \end{aligned}$$

This gives us the perceived game \mathbf{G}' in Matrix 2. For BoS the maximum incentive is $I_{\max} = S$ giving:

$$\begin{aligned} S'_j &= S - |S - I_j^*| & T'_j &= S - |T - I_j^*| \\ R'_j &= S - |R - I_j^*| & P'_j &= S - |P - I_j^*| \end{aligned}$$

This produces the perceived game \mathbf{G}' in Matrix 3. The next sections examine these perceived games when different values of I_j^* are assumed. We show that the games transform further into a series of new games with different NE depending on the value of I_j^* . There are numerous possible transformations of the game, but the remainder of this section focuses in theory on three cases of interest corresponding to individuals with strong power, achievement, and affiliation motivation. The simulations in section Results consider the intermediate cases as well.

Matrix 2. Perceived game \mathbf{G}' for PD, Leader, and Chicken.

		Agent A_2	
		D	C
Agent A_1	D	$T - P - I_1^* , T - P - I_2^* $	$T - T - I_1^* , T - S - I_2^* $
	C	$T - S - I_1^* , T - T - I_2^* $	$T - R - I_1^* , T - R - I_2^* $

Matrix 3. Perceived game \mathbf{G}' for Battle of the Sexes.

		Agent A_2	
		D	C
Agent A_1	D	$S - P - I_1^* , S - P - I_2^* $	$S - T - I_1^* , S - S - I_2^* $
	C	$S - S - I_1^* , S - T - I_2^* $	$S - R - I_1^* , S - R - I_2^* $

TRANSFORMING PRISONER'S DILEMMA

Using the PD game as an example, we can now consider how a game is transformed into new games, depending on the value of

I_j^* . Three cases are considered corresponding to individuals with strong power, achievement, and affiliation motivation.

Case 1 (Power): The first case examines a range of high optimally motivating incentives: $T > I_j^* > \frac{1}{2}(T + R)$. We consider this range "high" because I_j^* is closest to the maximum incentive T . This gives us the following transformation of the PD game using Matrix 2 and simplifying the absolute values using the assumption that $T > I_j^* > \frac{1}{2}(T + R) > R > P > S$:

$$T'_j = T - (T - I_j^*) = I_j^* \quad (3)$$

$$R'_j = T - (I_j^* - R) = T + R - I_j^* \quad (4)$$

$$P'_j = T - (I_j^* - P) = T + P - I_j^* \quad (5)$$

$$S'_j = T - (I_j^* - S) = T + S - I_j^* \quad (6)$$

Theorem 1. For a PD game \mathbf{G} with $T > R > P > S$, when $T > I_j^* > \frac{1}{2}(T + R)$ the perceived game \mathbf{G}' is still a valid PD with $T'_j > R'_j > P'_j > S'_j$.

Proof. If we assume $R'_j \geq T'_j$ then we have $T + R - I_j^* \geq I_j^*$ which simplifies to $\frac{1}{2}(T + R) \geq I_j^*$. This contradicts the assumption that $T > I_j^* > \frac{1}{2}(T + R)$ so it must be true that $T'_j > R'_j$. If we assume that $P'_j \geq R'_j$ then we have $T + P - I_j^* \geq T + R - I_j^*$ or $P \geq R$ which contradicts the definition of PD. Thus, it must be true that $R'_j > P'_j$. Likewise, if we assume that $S'_j \geq P'_j$ then we have $T + S - I_j^* \geq T + P - I_j^*$ which simplifies to $S \geq P$ which contradicts the definition of PD. Thus, it must be true that $P'_j > S'_j$. \square

Case 2 (Achievement): The second case examines a range of moderate optimally motivating incentives: $\frac{1}{2}(T + R) > I_j^* > R$. In other words, in this case I_j^* is closest to R . This gives us the same basic transformation of the PD game as in Case 1 (Equations 3–6), but now defines a different set of perceived game as follows:

Theorem 2. For a PD game \mathbf{G} with $T > R > P > S$, when $\frac{1}{2}(T + R) > I_j^* > R$ the perceived game \mathbf{G}' has $R'_j > T'_j$ and $P'_j > S'_j$.

Proof. If we assume $T'_j \geq R'_j$ then we have $I_j^* \geq T + R - I_j^*$ which simplifies to $I_j^* \geq \frac{1}{2}(T + R)$. This contradicts the assumption in this case that $\frac{1}{2}(T + R) > I_j^*$ so it must be true that $R'_j > T'_j$. If we assume that $S'_j \geq P'_j$ then we have $T + S - I_j^* \geq T + P - I_j^*$ which simplifies to $S \geq P$ which contradicts the definition of PD. Thus, it must be true that $P'_j > S'_j$. \square

Case 3 (Affiliation): The third case examines a range of low optimally motivating incentives: $\frac{1}{2}(P + S) > I_j^* > S$. We consider this range "low" because I_j^* is closest to S . This gives us the following transformation of the PD game using Matrix 2 and simplifying absolute values:

$$T'_j = T - (T - I_j^*) = I_j^*$$

$$\begin{aligned} R'_j &= T - (R - I_j^*) = T + I_j^* - R \\ P'_j &= T - (P - I_j^*) = T + I_j^* - P \\ S'_j &= T - (I_j^* - S) = T + S - I_j^* \end{aligned}$$

Theorem 3. For a PD game \mathbf{G} with $T > R > P > S$, when $\frac{1}{2}(P + S) > I_j^* > S$ the perceived game \mathbf{G}' has $S'_j > P'_j > R'_j > T'_j$.

Proof. If we assume $P'_j = S'_j$ then we have $T + I_j^* - P \geq T + S - I_j^*$ which simplifies to $I_j^* \geq \frac{1}{2}(P + S)$. This contradicts the assumption that $\frac{1}{2}(P + S) > I_j^*$. Thus, it must be true that $S'_j > P'_j$. If we assume $R'_j \geq P'_j$ then we have $T + I_j^* - R \geq T + I_j^* - P$ which simplifies to $P \geq R$. This contradicts the definition of PD. Thus, it must be true that $P'_j > R'_j$. Likewise, if we assume $T'_j \geq R'_j$ then we have $I_j^* \geq T + I_j^* - R$ which simplifies to $R \geq T$. This contradicts the definition of PD. Thus, it must be true that $R'_j > T'_j$ \square

The three cases above result in a number of different perceived games. Case 1 still results in a valid PD game, but in Case 2 and Case 3 the perceived games are new games. An example of the payoff structure of the new perceived game from Case 2 is visualized in **Figure 4A**. In this game $E_1(D) > E_1(C)$ for $P_2(C) = 0$ and $E_1(D) < E_1(C)$ for $P_2(C) = 1$. $E_1(D)$ and $E_1(C)$ intersect at:

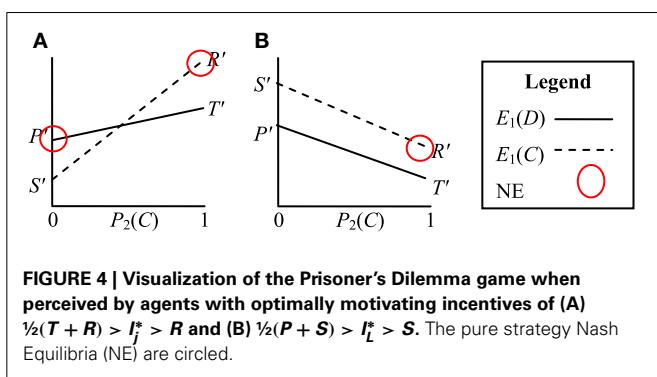
$$P_2(C) = \frac{P' - S'}{R' - S' - T' + P'} = M$$

There are now two pure NE and the strategy that emerges depends on the initial values of $P_1(C)$ and $P_2(C)$. If $P_1(C) + P_2(C) > 2M$ at $t = 0$ then the (C, C) equilibrium will emerge. Alternatively if $P_1(C) + P_2(C) < 2M$ at $t = 0$ then the (D, D) equilibrium will emerge.

In Case 3 the agents also do not perceive a PD game. The perceived game in this case is visualized in **Figure 4B**. In this game $E_1(C) > E_1(D)$ for all $P_2(C)$. The (C, C) strategy is now dominant, indicating that the agents will tend to evolve cooperative (C, C) strategies over time.

TRANSFORMING LEADER

We can follow the same process to construct perceived versions of Leader.



Case 1 (Power): The first case again examines a range of high optimally motivating incentives: $T > I_j^* > \frac{1}{2}(T + S)$. This gives us the same basic transformations in Equations 3–6, and the perceived game is still a Leader game.

Theorem 1. In a Leader game \mathbf{G} with $T > S > R > P$, when $T > I_j^* > \frac{1}{2}(T + S)$ the perceived game \mathbf{G}' is still a valid Leader game $T'_j > S'_j > R'_j > P'_j$.

Proof. If we assume $S'_j \geq T'_j$ then we have $T + S - I_j^* \geq I_j^*$ which simplifies to $\frac{1}{2}(T + S) \geq I_j^*$. This contradicts the assumption in this case that $T > I_j^* > \frac{1}{2}(T + S)$ so it must be true that $T'_j > S'_j$. If we assume that $R'_j \geq S'_j$ then we have $T + R - I_j^* \geq T + S - I_j^*$ which simplifies to $R \geq S$ which contradicts the definition of Leader. Thus, it must be true that $S'_j > R'_j$. Likewise, if we assume that $P'_j \geq R'_j$ then we have $T + P - I_j^* \geq T + R - I_j^*$ which simplifies to $P \geq R$ which contradicts the definition of Leader. Thus, it must be true that $R'_j > P'_j$ \square

Case 2 (Achievement): The second case examines a range of moderate-high optimally motivating incentive: $\frac{1}{2}(T + S) > I_j^* > S$. This also gives us the transformations in Equations 3–6, but the perceived game is no longer a Leader game. In fact, a number of interesting variations occur:

Lemma 1. In a Leader game \mathbf{G} with $T > S > R > P$, when $\frac{1}{2}(T + S) > I_j^* > S$ the perceived game \mathbf{G}' has $S'_j > T'_j$ and $R'_j > P'_j$.

Proof. If we assume $T'_j \geq S'_j$ then we have $I_j^* \geq T + S - I_j^*$ which simplifies to $I_j^* \geq \frac{1}{2}(T + S)$. This contradicts the assumption in this case that $\frac{1}{2}(T + S) > I_j^*$ so it must be true that $S'_j > T'_j$. If we assume that $P'_j \geq R'_j$ then we have $T + P - I_j^* \geq T + R - I_j^*$ which simplifies to $P \geq R$ which contradicts the definition of Leader. Thus, it must be true that $R'_j > P'_j$ \square

Theorem 2. In a Leader game \mathbf{G} with $T > S > R > P$, when $\frac{1}{2}(T + S) > I_j^* > S$ and $I_j^* > \frac{1}{2}(T + R)$ the perceived game \mathbf{G}' is a BoS game $S'_j > T'_j > R'_j > P'_j$

Proof. $S'_j > T'_j$ and $R'_j > P'_j$ by Lemma 3.2.2. $I_j^* > \frac{1}{2}(T + R)$ expands to $I_j^* > T + R - I_j^*$. Substitution of Equations 3–4 gives us $T'_j > R'_j$ \square

Theorem 3. In a Leader game \mathbf{G} with $T > S > R > P$, when $\frac{1}{2}(T + S) > I_j^* > S$ and $I_j^* < \frac{1}{2}(T + R)$ the perceived game \mathbf{G}' is $S'_j > R'_j > T'_j > P'_j$.

Proof. $S'_j > T'_j$ and $R'_j > P'_j$ by Lemma 3.2.2. $I_j^* < \frac{1}{2}(T + R)$ expands to $I_j^* < T + R - I_j^*$. Substitution of Equations 3–4 gives us $T'_j < R'_j$ \square

Case 3 (Affiliation): The third case examines a range of low optimally motivating incentives: $\frac{1}{2}(R+P) > I_j^* > P$. This gives us the

following transformation:

$$T'_j = T - [T - I_j^*] = I_j^* \quad (7)$$

$$R'_j = T - [R - I_j^*] = T + I_j^* - R \quad (8)$$

$$P'_j = T - [I_j^* - P] = T + P - I_j^* \quad (9)$$

$$S'_j = T - [S - I_j^*] = T + I_j^* - S \quad (10)$$

Theorem 4. In a Leader game \mathbf{G} with $T > S > R > P$, when $\frac{1}{2}(R + P) > I_j^* > P$ the perceived game \mathbf{G}' is $P'_j > R'_j > S'_j > T'_j$.

Proof. If we assume $R'_j \geq P'_j$ we have $T + I_j^* - R \geq T + P - I_j^*$ which simplifies to $I_j^* \geq 1/2(R + P)$ which contradicts the assumption that $1/2(R + P) > I_j^*$. If we assume $S'_j \geq R'_j$ we have $T + I_j^* - S \geq T + I_j^* - R$ or $R \geq S$ which contradicts the definition of Leader. Thus, it must be true that $R'_j > S'_j$. Likewise if we assume $T'_j \geq S'_j$ we have $I_j^* \geq T + I_j^* - S$ or $S \geq T$ which contradicts the definition of Leader. Thus, it must be true that $S'_j > T'_j$. \square

TRANSFORMING CHICKEN

We can follow the same process again to construct the perceived versions of Chicken. Proofs are omitted for brevity.

Case 1 (Power): The first case again assumes a high optimally motivating incentive: $T > I_j^* > 1/2(T + R)$. This gives us the transformation in Equations 3–6, and the perceived game is a Chicken game:

Theorem 1. For a Chicken game \mathbf{G} with $T > R > S > P$, when $T > I_j^* > 1/2(T + R)$ the perceived game \mathbf{G}' is still a valid Chicken game $T'_j > R'_j > S'_j > P'_j$.

Proof. Omitted. \square

Case 2 (Achievement): The second case again assumes a moderate-high optimally motivating incentive: $1/2(T + R) > I_j^* > R$. This also gives us the transformation in Equations 3–6, but the perceived game is no longer a Chicken game:

Theorem 2. For a Chicken game \mathbf{G} with $T > R > S > P$, when $\frac{1}{2}(T + R) > I_j^* > R$ the perceived game \mathbf{G}' has $R'_j > T'_j$ and $S'_j > P'_j$.

Proof. Omitted. \square

Case 3 (Affiliation): The third case again assumes a low optimally motivating incentive: $\frac{1}{2}(S + P) > I_j^* > P$. This gives us the transformations in Equations 7–10.

Theorem 3. For a Chicken game \mathbf{G} with $T > R > S > P$, when $\frac{1}{2}(S + P) > I_j^* > P$ the perceived game \mathbf{G}' is $P'_j > S'_j > R'_j > T'_j$.

Proof. Omitted. \square

TRANSFORMING BATTLE OF THE SEXES

Finally, we can follow the process above to construct the perceived versions of BoS.

Case 1 (Power): The first case again assumes a high optimally motivating incentive: $S > I_j^* > \frac{1}{2}(T + S)$. This gives us the following transformation of the BoS game:

$$T'_j = S - (I_j^* - T) = S + T - I_j^* \quad (11)$$

$$R'_j = S - (I_j^* - R) = S + R - I_j^* \quad (12)$$

$$P'_j = S - (I_j^* - P) = S + P - I_j^* \quad (13)$$

$$S'_j = S - (S - I_j^*) = I_j^* \quad (14)$$

Theorem 1. For a BoS game \mathbf{G} with $S > T > R > P$, when $S > I_j^* > \frac{1}{2}(T + S)$ the perceived game \mathbf{G}' is still a valid BoS game $S'_j > T'_j > R'_j > P'_j$.

Proof. Omitted. \square

Case 2 (Achievement): The second case again assumes a moderate-high optimally motivating incentive: $\frac{1}{2}(T + S) > I_j^* > T$. This gives us the transformation of the BoS game in Equations 11–14, but the perceived game is no longer a BoS.

Lemma 1. For a BoS game \mathbf{G} with $S > T > R > P$, when $\frac{1}{2}(T + S) > I_j^* > T$ the perceived game \mathbf{G}' has $T'_j > S'_j$ and $R'_j > P'_j$.

Proof. If we assume $S'_j \geq T'_j$ then we have $I_j^* \geq S + T - I_j^*$ which simplifies to $I_j^* \geq \frac{1}{2}(T + S)$ which contradicts the assumption that $\frac{1}{2}(T + S) > I_j^*$. Thus, it must be true that $S'_j > T'_j$. If we assume $P'_j \geq R'_j$ then we have $S + P - I_j^* \geq S + R - I_j^*$ which simplifies to $P \geq R$ which contradicts the definition of BoS. Thus, it must be true that $R'_j > P'_j$. \square

Theorem 2. For a BoS game \mathbf{G} with $S > T > R > P$, when $\frac{1}{2}(T + S) > I_j^* > T$ and $I_j^* > \frac{1}{2}(S + R)$ the perceived game \mathbf{G}' is a Leader game $T'_j > S'_j > R'_j > P'_j$.

Proof. $T'_j > S'_j$ and $R'_j > P'_j$ by Lemma 3.4.2. $I_j^* > \frac{1}{2}(S + R)$ expands to $I_j^* > S + R - I_j^*$. Substitution of Equations 14 and 12 gives us $S'_j > R'_j$. \square

Theorem 3. For a BoS game \mathbf{G} with $S > T > R > P$, when $\frac{1}{2}(T + S) > I_j^* > T$ and $I_j^* < \frac{1}{2}(S + R)$ the perceived game \mathbf{G}' is a Chicken game $T'_j > R'_j > S'_j > P'_j$.

Proof. $T'_j > S'_j$ and $R'_j > P'_j$ by Lemma 3.4.2. $I_j^* < \frac{1}{2}(S + R)$ expands to $I_j^* < S + R - I_j^*$. Substitution of Equations 14 and 12 gives us $S'_j < R'_j$. \square

Case 3 (Affiliation): The third case again assumes a low optimally motivating incentive: $\frac{1}{2}(R + P) > I_j^* > P$. This gives us the

following transformation of the BoS game:

$$\begin{aligned} T'_j &= S - (T - I_j^*) = S + I_j^* - T \\ R'_j &= S - (R - I_j^*) = S + I_j^* - R \\ P'_j &= S - (I_j^* - P) = S + P - I_j^* \\ S'_j &= S - (S - I_j^*) = I_j^* \end{aligned}$$

Theorem 4. For a BoS game \mathbf{G} with $S > T > R > P$, when $\frac{1}{2}(R + P) > I_j^* > P$ the perceived game \mathbf{G}' is $P'_j > R'_j > T'_j > S'_j$.

Proof. If we assume $R'_j \geq P'_j$ then we have $S + I_j^* - R \geq S + P - I_j^*$ or $I_j^* \geq \frac{1}{2}(R + P)$ which contradicts the assumption that $\frac{1}{2}(R + P) > I_j^*$. Thus, it must be true that $P'_j > R'_j$. If we assume that $T'_j \geq R'_j$ then we have $S + I_j^* - T \geq S + I_j^* - R$ or $R \geq T$ which contradicts the definition of BoS. Thus, it must be true that $R'_j > T'_j$. Likewise, if we assume that $S'_j \geq T'_j$ then we have $I_j^* \geq S + I_j^* - T$ or $T \geq S$ which contradicts the definition of BoS. Thus, it must be true that $T'_j > S'_j$ \square

RESULTS

This section presents simulations of the each of the four games studied in section Materials and Methods played by agents with optimally motivating incentives conforming to the three cases studied, as well as the intermediate cases not studied above. We use two-population replicator dynamics to model cultural learning when mixed-motive games are played iteratively. We demonstrate that individuals with different optimally motivating incentives may adopt different strategies when playing a particular game, or may learn at different rates. We also discuss how the NE of the transformed games reflects a number of results from human experiments that are not well-modeled by the NE of the original game.

PRISONERS' DILEMMA

Figures 5, 6 use the two population replicator dynamics in Equations 1 and 2 to simulate one hundred pairs of agents (A_1 and A_2) playing the iterated PD (IPD¹) game:

$$\mathbf{G} = \begin{bmatrix} 2 & 4 \\ 1 & 3 \end{bmatrix}$$

The initial probabilities p_C (for agents A_1) and q_C (for agents A_2) are randomized and the agent pairs learn while playing thirty consecutive games. A range of [1, 4] is assumed for incentive. The lines in **Figure 5** trace the learned values of p_C and q_C over time. In **Figure 5** all agents have a “high” optimally motivating incentive $I_1^* = I_2^* = 4.0$, representing power-motivated individuals. We see that the perceived games are identical to the original game, ie: $\mathbf{G}'_1 = \mathbf{G}'_2 = \mathbf{G}$ and all agent pairs tend to converge on the (D, D) equilibrium over time.

In **Figure 6** the agents share progressively lower values of I_1^* and I_2^* , ranging from $I_1^* = I_2^* = 3.8$ in **Figure 6A** to $I_1^* = I_2^* = 1.0$ in **Figure 6O**. **Figures 6A,B** show Case 1 games in which the (D, D) outcome emerges as the equilibrium as predicted by

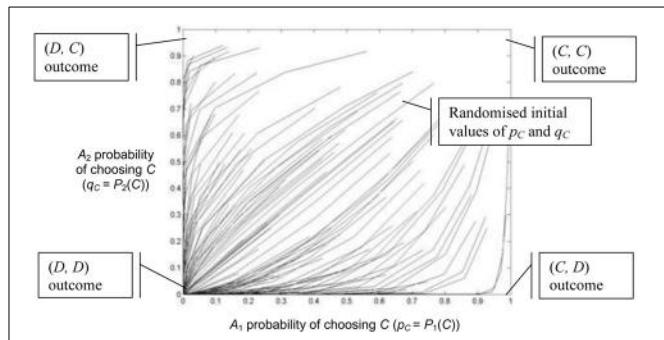


FIGURE 5 | Simulation of one hundred pairs of agents playing thirty iterations of the Prisoner's Dilemma game. All agents have $I_j^* = 4.0$, but initial values of p_C and q_C are randomized.

Theorem 2.1.1. These agents still perceive a PD game. In contrast, **Figures 6C,D** show Case 2 games in which some agents converge on the (C, C) equilibrium and some on the (D, D) equilibrium, as predicted by Theorem 2.1.2. The equilibrium approached by the agent pairs in this case depends on their initial values of p_C and q_C . In **Figures 6E–L** the (C, C) outcome becomes more frequent as the values of I_1^* and I_2^* decrease. **Figures 6M,N** shows Case 3 games in which all agents converge on the (C, C) equilibrium as predicted by Theorem 2.1.3.

In general, these results support the idea proposed by Johnson et al. (2002), that individual variation means that true PD scenarios occur relatively infrequently in nature. Johnson et al. (2002) show that if there is variance in perception of twice the payoff interval in a linear PD game (a game in which the intervals between T , R , S , and P are the same) then only 15.8% remain valid PD games. Our transformations show that a true PD scenario will only occur if both agents have optimally motivating incentives that fall in the range $T > I^* > \frac{1}{2}(T + R)$. If we assume I^* can only fall within the range $T \geq I^* \geq S$, the fraction v of valid PD games will be:

$$v = \frac{T - \frac{1}{2}(T + R)}{T - S} = \frac{T - R}{2(T - S)}$$

In a linear PD game $3(T - R) = (T - S)$ so $v = 1/6 = 16.6\%$ if we assume a uniform distribution of optimally motivating incentives. This is, qualitatively speaking, similar to the result proposed by Johnson et al. (2002), and offers support for our methodology for modeling differences in motivations.

Case 1 and Case 2 also provide computational insight into some of the findings reported by Terhune (1968). Terhune observed pairs of humans classified as either power, affiliation and achievement motivated playing single-shot and iterative PD games in controlled conditions. One of these experiments observed the influence of the first trial outcome on different types of people. He found that if the first outcome was (C, C) , pairs of achievement motivated individuals had the highest subsequent proportion of (C, C) outcomes (46.8%). In contrast, power motivated individuals had (C, C) outcomes only 9.4% of the time after a (C, C) outcome on the first trial. In other words people with different motives respond differently to the

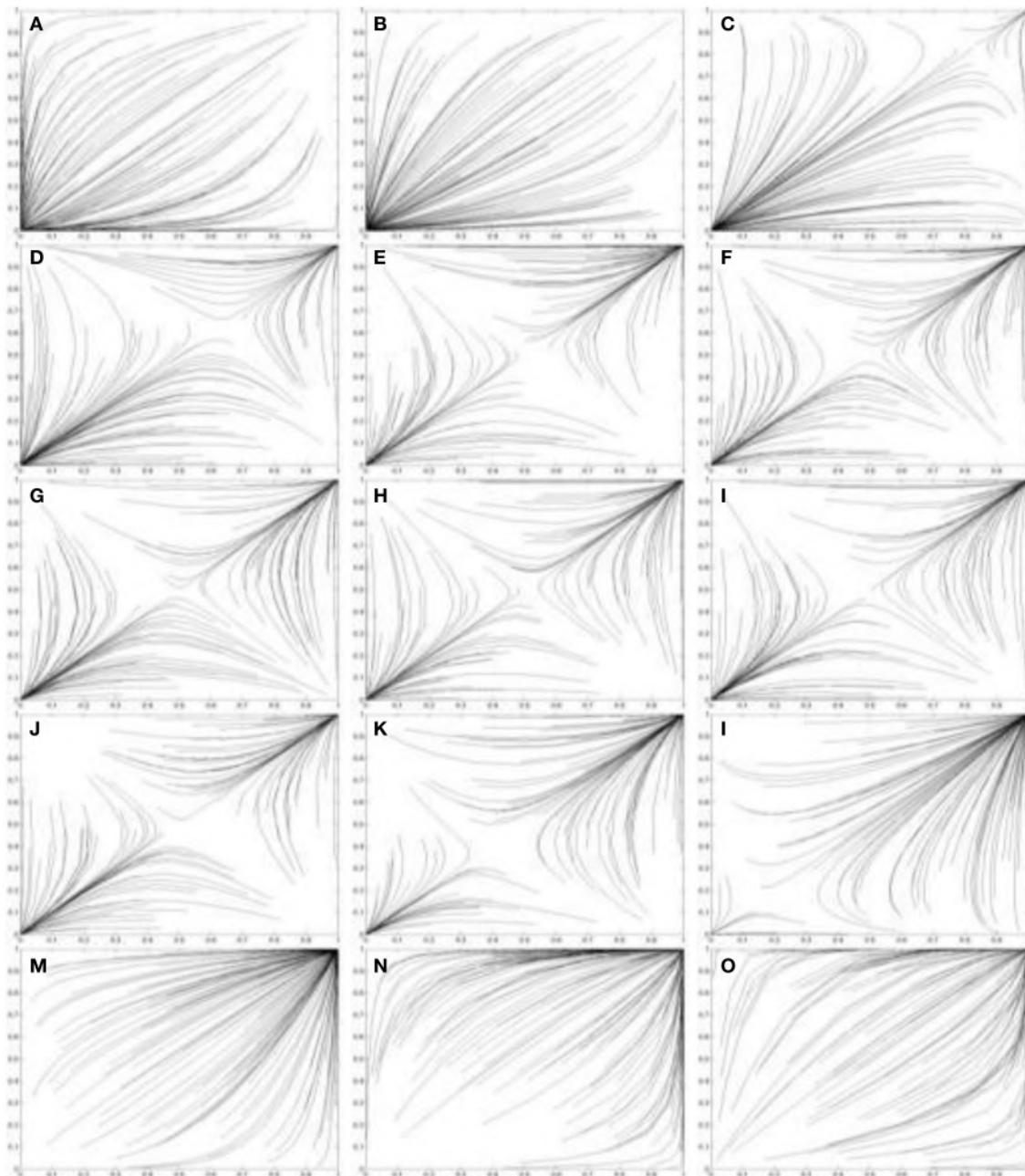


FIGURE 6 | Simulations of one hundred pairs of agents playing thirty iterations of the Prisoner's Dilemma game. Agents share different values of I_j^* in each simulation. (A) $I_j^* = 3.8$; (B) $I_j^* = 3.6$; (C) $I_j^* = 3.4$; (D) $I_j^* = 3.2$;

(E) $I_j^* = 3.0$; (F) $I_j^* = 2.8$; (G) $I_j^* = 2.6$; (H) $I_j^* = 2.4$; (I) $I_j^* = 2.2$; (J) $I_j^* = 2.0$; (K) $I_j^* = 1.8$; (L) $I_j^* = 1.6$; (M) $I_j^* = 1.4$; (N) $I_j^* = 1.2$; (O) $I_j^* = 1.0$. Initial values of p_C and q_C are randomized. See Figure 5 for legend.

same experience (in this case the first trial outcome). The results above suggest that this can be captured computationally using our model by using high values of I^* for power motivated individuals, so that they tend to perceive a Case 1 game and lower values of I^* for achievement motivated individuals, so that they tend to perceive a Case 2 game. A further discussion of this avenue for future work is made in section Human-Computer Interaction.

The Case 3 result is perhaps less instructive from a human modeling perspective, but is still useful from an artificial systems perspective. If we wish to design agents that will cooperate when faced with PD situations, then we can use agents with low optimally motivating incentives in the range $\frac{1}{2}(P + S) > I_1^* > S$. These agents perceive a game with a dominant (C, C) strategy and will thus tend to evolve cooperative strategies over time. Likewise, if we wish to model “martyrs” then an agent A_1 with

$\frac{1}{2}(P + S) > I_1^* > S$ will be a martyr (C chooser) when playing an agent A_2 with $T > I_2^* > \frac{1}{2}(T + R)$. This type of personality modeling has application to areas such as believable non-player characters (NPCs) in computer games.

LEADER

If we consider Case 1(power-motivated) agents playing the leader game, we see that $E_1(C) > E_1(D)$ for $P_2(C) = 0$ and $E_1(D) > E_1(C)$ for $P_2(C) = 1$. $E_1(C)$ and $E_1(D)$ intersect at the point:

$$P_2(C) = \frac{S - P}{2I^* + S - P - T - R}$$

Now, suppose we have two pairs of players. The first pair of players have optimally motivating incentives $I_1^* = I_2^* = I_j^*$. The second pair of players have optimally motivating incentives $I_1^* = I_2^* = I_k^*$ such that $I_j^* > I_k^*$. Substitution gives us

$$\frac{S - P}{2I_j^* + S - P - T - R} < \frac{S - P}{2I_k^* + S - P - T - R}$$

That is, $P_j(C) < P_k(C)$. In other words the probability of conceding right of way increases in games between players with weaker power motivation, although the equilibria are still at (C, D) and (D, C) as indicated by Theorem 2.2.1. This phenomenon is evident in the simulations in **Figure 7**. **Figure 7** uses the two population replicator dynamics in Equations 1 and 2 to simulate one hundred pairs of learning agents (A_1 and A_2) playing the Leader game:

$$G = \begin{bmatrix} 1 & 4 \\ 3 & 2 \end{bmatrix}$$

The Case 1 simulations are shown in **Figures 7A,B** and the trend to concede is evident in the progressively less direct paths the agent's take to the equilibria. As I_j^* is further decreased in Case 2 (achievement motivated agents), two types of perceived games occur. Either the game is perceived as a BoS game (Theorem 2.2.3), or as a game with a dominant (C, C) strategy (Theorem 2.2.4).

The Leader game is perceived as a BoS game when $\frac{1}{2}(T + S) > I_j^* > S$ and $I_j^* = \frac{1}{2}(T + R)$. The payoff structure for a BoS game is visualized in **Figure 2C**. **Figures 7C,D** simulates the behavior of agents that perceive a Leader game as a BoS game. The paths taken to the (C, D) and (D, C) equilibria by these agents are quite indirect as both are initially motivated to concede right of way by their perception of leadership as an act of sacrifice. Leader-follower behavior [(C, D) or (D, C)] does emerge, but it does so more slowly than for agents with high values of I_j^* because leadership is now perceived as an act of sacrifice.

Figures 7E–J shows simulations of games between agents with $S > I_j^* > R$. These agents perceive games of the form $S'_j > R'_j > T'_j > P'_j$ with dominant (C, C) strategies. As a result, leadership behavior does not emerge as an equilibrium as the agents always concede right of way. In Case 3(affiliation motivated agents) there are two pure equilibria in the perceived game: (D, D) and (C, C) .

The Case 3 payoff structure is simulated in **Figures 7M,N**. The emergent equilibrium strategy for any pair of agents depends on the initial values of $P_1(C)$ and $P_2(C)$. If $P_1(C) + P_2(C) > 2M$ at $t = 0$ then the (C, C) equilibrium will occur over time. Alternatively if $P_1(C) + P_2(C) < 2M$ at $t = 0$ then the (D, D) equilibrium will occur over time. These pure strategy equilibria preclude the emergence of leader-follower behavior and result, instead, in collisions (both players driving) or procrastination (both players conceding right of way). Thus, to achieve leaders and followers agents with high values of I^* are required.

CHICKEN

In the chicken game, Case 1(power-motivated) agents also perceive a valid Chicken game resulting in the emergence of an “exploiter” agent. However, with a small reduction in I_j^* Case 2 (achievement motivated) agents perceive a transformed game in which the more cautious (C, C) strategy is dominant (Theorem 2.3.2). This is, in fact, the most common perceived game, covering $\frac{1}{2}(T + R) > I_j^* > \frac{1}{2}(S + P)$. This can be thought of as reflecting the real-world reluctance to engage in a game of Chicken, which is in principle the same as playing and choosing C (Colman, 1982).

The prevalence of the perceived dominant (C, C) strategy is evidenced in the simulations in **Figure 8**. **Figure 8** uses the two population replicator dynamics in Equations 1 and 2 to simulate one hundred pairs of learning agents (A_1 and A_2) playing the Chicken game:

$$G = \begin{bmatrix} 1 & 4 \\ 2 & 3 \end{bmatrix}$$

Figures 8C–L all show agents approaching the (C, C) equilibrium. One other case does exist (Case 3) in which the perceived game has two pure NE: (D, D) and (C, C) . The emergent equilibrium for two agents depends on the initial values of $P_1(C)$ and $P_2(C)$. If $P_1(C) + P_2(C) > 2M$ at $t = 0$ then the (C, C) equilibrium will occur over time. Alternatively if $P_1(C) + P_2(C) < 2M$ at $t = 0$ then the (D, D) equilibrium will occur over time. These pure strategy equilibria result in either certain collision (both players driving on) or mutually cautious behavior (both players swerving to avoid a collision). Examples of Case 3 agents interacting are shown in **Figures 7M,N**.

Comparison of Case 1 and Case 3 demonstrates how the same outcome may result from different motives. In Case 1 the (D, D) outcome results from a preference for high incentives. In Case 3 the (D, D) outcome results from a preference for low incentives to avoid conflict. The strategy clearly backfires, but this sort of trend has been observed in a general sense in humans. Individuals with high affiliation motivation have been observed to underperform their achievement motivated colleagues precisely because their desire to avoid conflict situations often means they also miss opportunities to cooperate (Heckhausen and Heckhausen, 2008).

BATTLE OF THE SEXES

If we consider Case 1 (power-motivated) agents playing BoS, we see that $E_1(C) > E_1(D)$ for $P_2(C) = 0$ and $E_1(D) > E_1(C)$ for

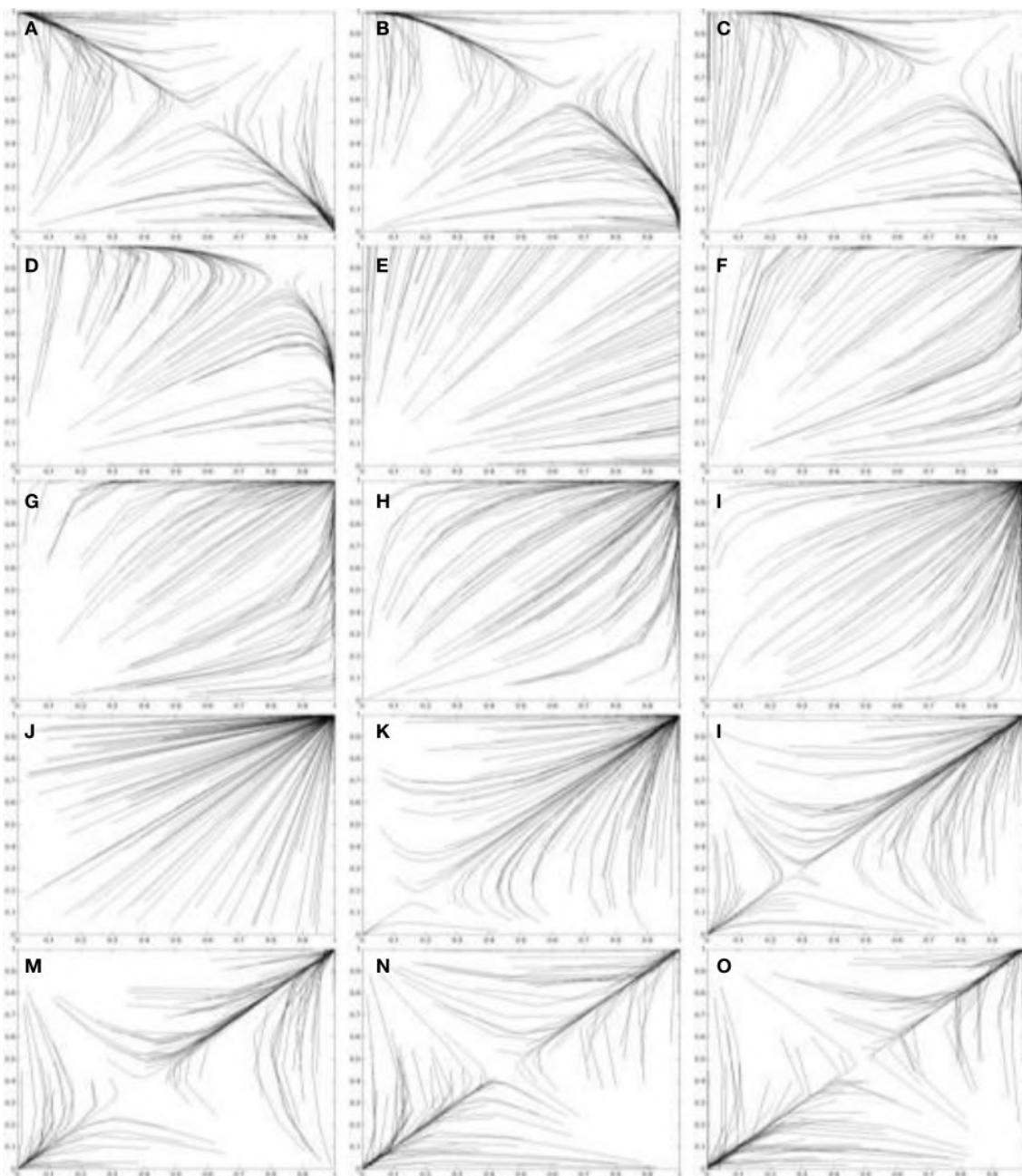


FIGURE 7 | Simulations of one hundred pairs of agents playing thirty iterations of the Leader game. Agents share different values of I_j^* in each simulation. (A) $I_j^* = 3.8$; (B) $I_j^* = 3.6$; (C) $I_j^* = 3.4$; (D) $I_j^* = 3.2$; (E) $I_j^* = 3.0$; (F) $I_j^* = 2.8$; (G) $I_j^* = 2.6$; (H) $I_j^* = 2.4$; (I) $I_j^* = 2.2$; (J) $I_j^* = 2.0$; (K) $I_j^* = 1.8$; (L) $I_j^* = 1.6$; (M) $I_j^* = 1.4$; (N) $I_j^* = 1.2$; (O) $I_j^* = 1.0$. Initial values of p_C and q_C are randomized. See Figure 5 for legend.

$P_2(C) = 1$. $E_1(C)$ and $E_1(D)$ intersect at the point:

$$P_2(C) = \frac{2I^* - S - P}{2I^* - S - P + T - R}$$

Now, suppose we have two pairs of learning agents playing a BoS game. The first pair of agents has optimally motivating incentives $I_1^* = I_2^* = I_j^*$. The second pair has optimally

motivating incentives $I_1^* = I_2^* = I_k^*$ such that $I_j^* < I_k^*$. This implies $P_j(C) < P_k(C)$ as the $(T - R)$ term in the denominator becomes increasingly significant as I^* decreases. In other words, the probability of choosing C decreases in agents with lower values of I^* as they begin to perceive the D choice as a desirable act of leadership rather than as a less desirable act of sacrifice. This is evident in the simulations in Figure 9. Figure 9 uses the two population replicator dynamics in Equations 1 and 2 to simulate one

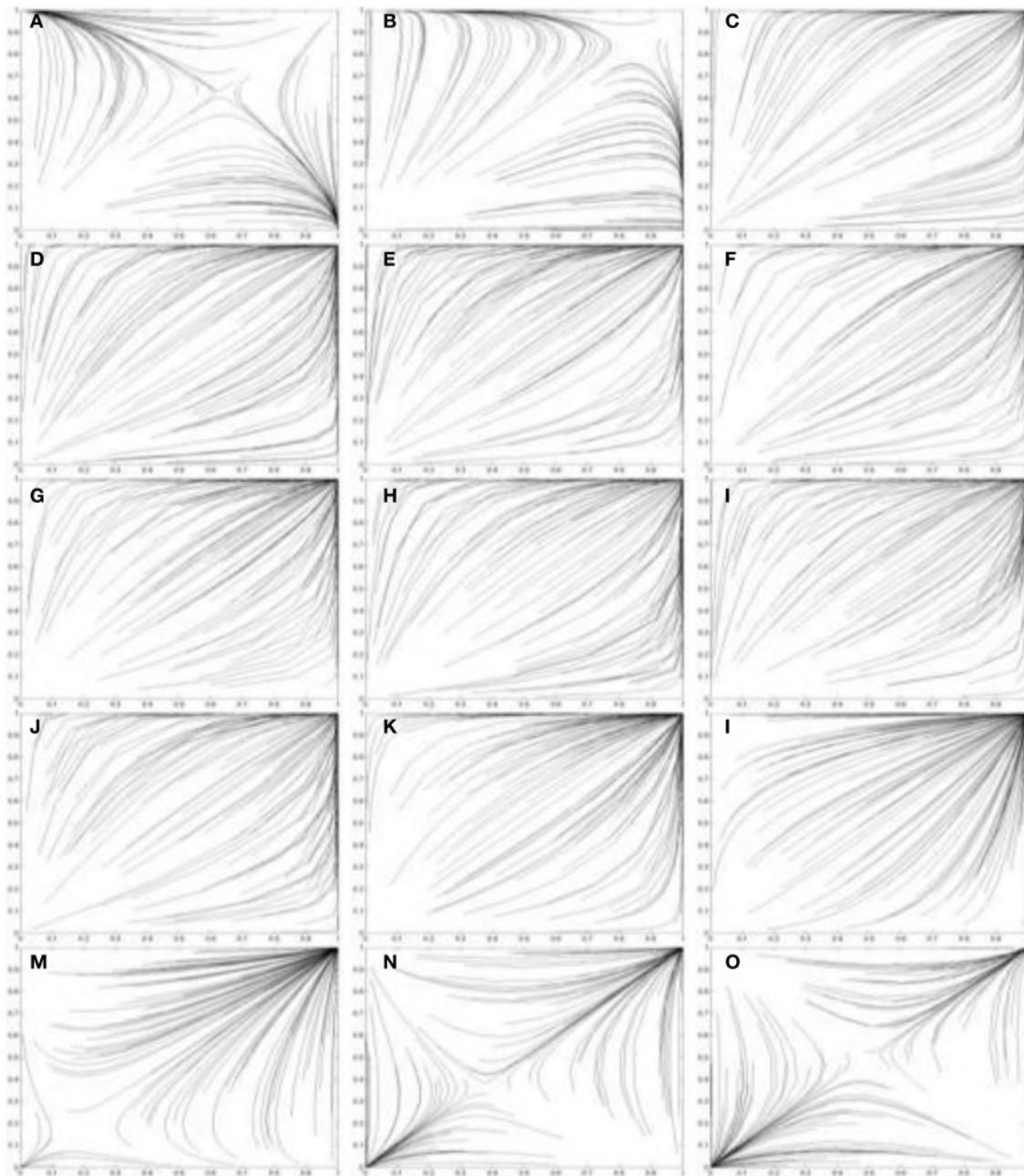


FIGURE 8 | Simulations of one hundred pairs of agents playing thirty iterations of the Chicken game. Agents share different values of I_j^* in each simulation. (A) $I_j^* = 3.8$; (B) $I_j^* = 3.6$; (C) $I_j^* = 3.4$; (D) $I_j^* = 3.2$; (E) $I_j^* = 3.0$; (F) $I_j^* = 2.8$; (G) $I_j^* = 2.6$; (H) $I_j^* = 2.4$; (I) $I_j^* = 2.2$; (J) $I_j^* = 2.0$; (K) $I_j^* = 1.8$; (L) $I_j^* = 1.6$; (M) $I_j^* = 1.4$; (N) $I_j^* = 1.2$; (O) $I_j^* = 1.0$. Initial values of p_C and q_C are randomized. See Figure 5 for legend.

hundred pairs of agents (A_1 and A_2) playing the BoS game:

$$\mathbf{G} = \begin{bmatrix} 1 & 3 \\ 4 & 2 \end{bmatrix}$$

Figures 9A,B show Case 1 simulations while **Figures 9C,D** show Case 2 simulations in which the learning agents perceive a Leader game (Theorem 2.4.3) rather than the original BoS game.

Progressively more direct trajectories towards the (C, D) and (D, C) outcomes are evident in these simulations as I_j^* decreases.

Figures 9E–G show simulations in which the agents perceive a Chicken game rather than a BoS game. This is followed by another change in perception in **Figures 9H,L**. In these simulations, and in the Case 3 games in **Figures 9M,N** the perceived games have two pure NE: (D, D) and (C, C) . The strategy chosen by the agents depends on the initial values of p_C and q_C . These

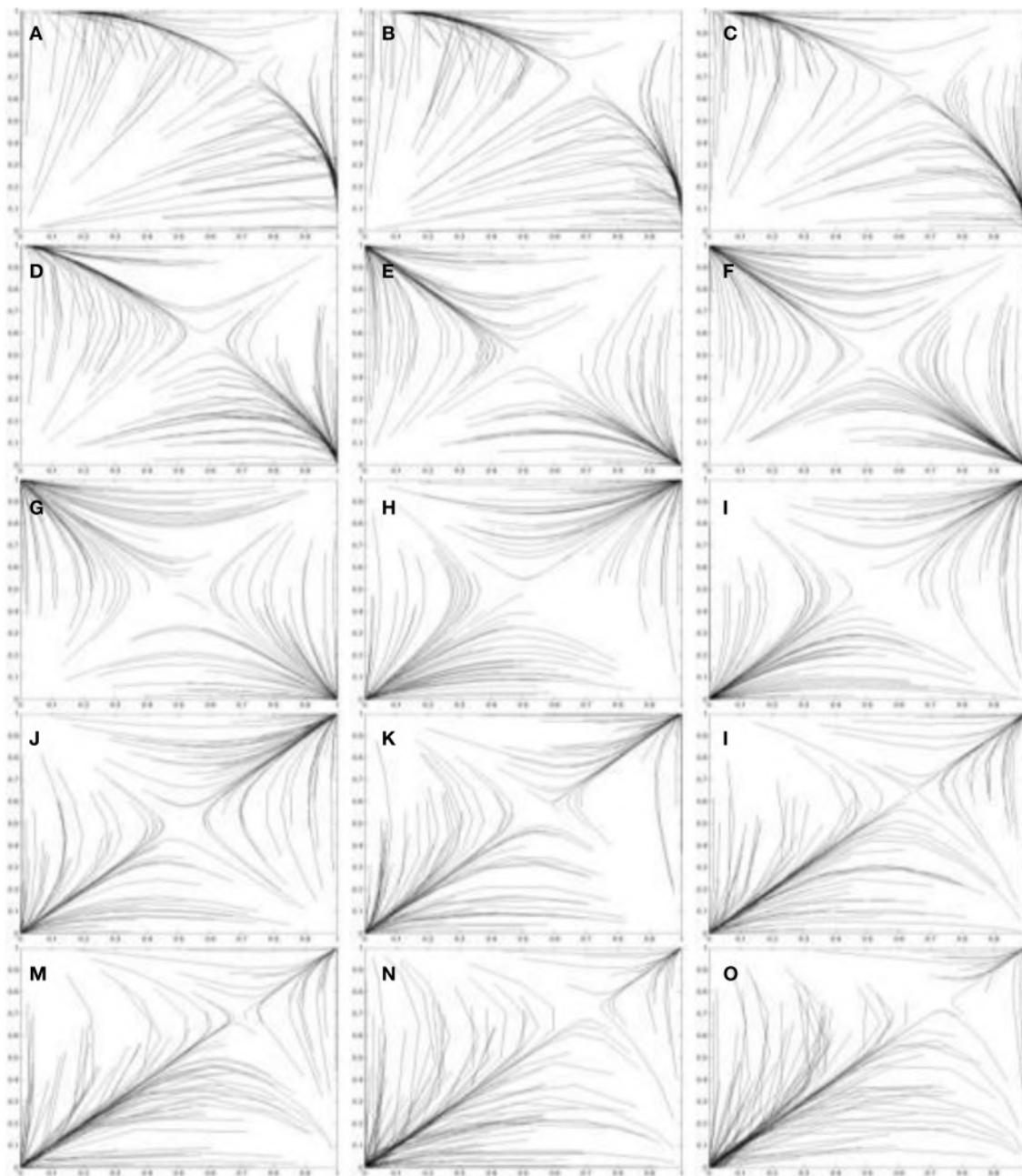


FIGURE 9 | Simulations of one hundred pairs of agents playing thirty iterations of the Battle-of-the-Sexes game. Agents share different values of I_j^* in each simulation. **(A)** $I_j^* = 3.8$; **(B)** $I_j^* = 3.6$; **(C)** $I_j^* = 3.4$; **(D)** $I_j^* = 3.2$; **(E)** $I_j^* = 3.0$; **(F)** $I_j^* = 2.8$; **(G)** $I_j^* = 2.6$; **(H)** $I_j^* = 2.4$; **(I)** $I_j^* = 2.2$; **(J)** $I_j^* = 2.0$; **(K)** $I_j^* = 1.8$; **(L)** $I_j^* = 1.6$; **(M)** $I_j^* = 1.4$; **(N)** $I_j^* = 1.2$; **(O)** $I_j^* = 1.0$. Initial values of p_C and q_C are randomized. See **Figure 5** for legend.

pure strategy equilibria result in both players attending entertainment alone. For the best outcome to emerge, either a “hero,” a “leader,” or a “chicken” personality is required.

STRATEGIC INTERACTIONS BETWEEN AGENTS WITH DIFFERENT MOTIVES

The simulations so far consider pairs of agents with the same optimally motivating incentives. However, it is also possible to

simulate the outcomes when pairs of learning agents with different optimally motivating incentives interact. **Figures 10A–D** simulates such pairs of agents playing each of the four games, PD, Leader, Chicken, and BoS, respectively. In each pair, one agent A_1 has a high optimally motivating incentive $I_1^* = 3.9$ and the other A_2 has a low optimally motivating incentive $I_1^* = 1.1$.

The results in **Figure 10** show that agents with high optimally motivating incentive tend to be the “exploiters” in PD and

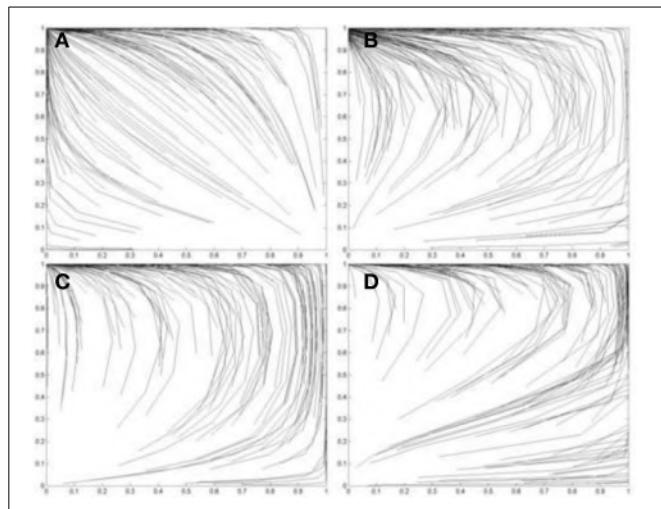


FIGURE 10 | Simulations of one hundred pairs of agents playing thirty iterations of (A) the Prisoner’s Dilemma game; (B) the Leader game; (C) the Chicken game; and (D) the Battle-of-the-Sexes game. In each simulation, one agent in each pair has $I_1^* = 3.9$ and the other has $I_2^* = 1.1$. Initial values of p_C and q_C are randomized. See Figure 5 for legend.

Chicken games, the “leaders” in a Leader game, and the “heroes” in a BoS game. In contrast, agents with low optimally motivating incentive (less than the average of the lowest two payoffs of a game) tend to be the “martyrs” in a PD game, the “followers” in a Leader game, the “chickens” in a Chicken game and the “selfish” in a BoS game.

DISCUSSION

In this paper we have represented agents with an optimally motivating incentive that influences the way they perceive the payoffs in strategic interactions. By using two-by-two mixed-motive games to represent different kinds of strategic interactions, we have shown that agents with different optimally motivating incentives perceive the original game differently. In many cases the perceived games have different equilibrium points to the original game. We can draw a number of general conclusions about the perceptions of agents with different optimally motivating incentives:

- Agents with high optimally motivating incentive (greater than the average of the highest two payoffs of a game) perceive a game that still conforms to the conditions defining the original game. For example, an agent with high optimally motivating incentive playing a PD game will still perceive a valid PD game and so on.
- Agents with moderate or lower optimally motivating incentive perceive new games that do not conform to the conditions defining the original game. This changes the NE and the behavior of the agents over time.

When agents with different optimally motivating incentives interact:

- Agents with high optimally motivating incentive will tend to be the “exploiters” in PD and Chicken games, the

“leaders” in a Leader game, and the “heroes” in a BoS game.

- Agents with low optimally motivating incentive (less than the average of the lowest two payoffs of a game) will tend to be the “martyrs” in a PD game, the “followers” in a Leader game, the “chickens” in a Chicken game and the “selfish” in a BoS game.

The concept of optimally motivating incentive thus provides an approach to building artificial agents with different personalities using motivation. Personality in this case is expressed through behavior. For example, using the language of Colman (1982), agents in the simulations in section Results can be interpreted as demonstrating behavioral characteristics such as “aggression,” “leadership,” “heroism,” “martyrdom,” and “caution.” This suggests a number of possible applications including the design of more believable agents, human-computer interaction and simulation of human decision-making. These are discussed in the following sub-sections.

BELIEVABLE AGENTS

Agents with distinguishable personalities have applications in areas such as animated entertainment where believable agents increase the sense of immersion in a virtual environment. According to Loyall (1997), believable agents should “allow people to not just watch, but also interact with... powerful, personality-rich characters.” The work in this paper specifically explores the role of intrinsic motivation for artificial agents engaged in social interactions. While the experiments in this paper are abstracted to the decision-making level, it is feasible to imagine an extension of this work in which this decision making controls the animated behaviour of a virtual character.

Some existing work has studied self-motivated behavior such as curiosity and novelty-seeking in NPCs in computer games (Merrick and Maher, 2009). Merrick and Maher (2009) demonstrate that intrinsically motivated reinforcement learning agents can learn in open-ended environments by generating goals in response to their experiences. The simulations in this paper combined optimally motivating incentive with learning using replicator dynamics, to complement the analytical description of each game transformation. However, in future it is feasible that motive profiles may be combined with learning algorithms that learn from actual interaction and experimentation with their environment during strategic interactions. Reinforcement learning variants such as frequency adjusted Q-learning (Kaisers and Tuyls, 2010) have been specifically developed for such multi-agent systems and suggest a starting point for such work. This would permit a wider range of motives to be used in NPCs. It would also extend existing work with intrinsically motivated NPCs from scenarios in which individual agents interact with their environment to scenarios in which multiple intrinsically motivated agents interact with each other.

HUMAN-COMPUTER INTERACTION

Just as the study of computational models of motivation lies at the intersection of computer science and cognitive science, another area of future work lies at the boundary where computer and human interact. In particular, computers are increasingly applied to problems that require them to develop beliefs

about the motives and intentions of the humans with whom they interact. Maher et al. (2007) for example, propose “*curious places*” in which a building is an “immobile robot” with sensors and actuators permitting it to monitor and control the built environment. The aim of the immobile robot is to intervene proactively on behalf of the human and modify the environment in a manner that supports the human’s goals. In order to do this, it must first identify those goals.

The framework in this paper can be conceived as a foundation for agents to simulate and reason about the decision-making of other agents or humans. As discussed in section Mixed-Motive Games, the four games studied in this paper represent abstractions of real-world interaction scenarios. A robot equipped with appropriate sensors might monitor the behavior of a given human in such scenarios and deduce their motive profile from their behavior. By engaging in such “autonomous mental simulation” of the intrinsically motivated reasoning of another, such an agent may ultimately be better equipped to estimate and support the goals of humans.

SIMULATION OF HUMAN DECISION-MAKING

The theories presented in this paper provide a starting point for developing populations of agents that can reproduce certain aspects of human decision-making during strategic interactions. Merrick and Shafi (2011) showed that it is possible to calibrate power, achievement and affiliation motivated agents such that

REFERENCES

- Atkinson, J. W. (1957). Motivational determinants of risk-taking behavior. *Psychol. Rev.* 64, 359–372. doi: 10.1037/h0043445
- Atkinson, J. W., and Litwin, G. H. (1960). Achievement motive and test anxiety conceived as motive to approach success and motive to avoid failure. *J. Abnorm. Soc. Psychol.* 60, 52–63. doi: 10.1037/h0041119
- Atkinson, J. W., and Raynor, J. O. (1974). *Motivation and Achievement*. Washington, DC: V. H. Winston.
- Borgers, T., and Sarin, R. (1997). Learning through reinforcement and replicator dynamics. *J. Econ. Theory* 77, 1–14. doi: 10.1006/jeth.1997.2319
- Camerer, C. (2003). *Behavioral Game Theory: Experiments in Strategic Interaction*. New Jersey, NJ: Princeton University Press.
- Camerer, C. (2004). Behavioral game theory: predicting human behavior in strategic situations,” in *Advances in Behavioural Economics*, eds C. Camerer, G. Loewenstein, and M. Rabin. (New York, NY: Princeton University Press), 374–392.
- Claus, C., and Boutilier, C. (1998). “The dynamics of reinforcement learning in cooperative multiagent systems,” in *The National Conference on Artificial Intelligence (AAAI 1998)* (Madison, WI).
- Colman, A. (1982). “Game theory and experimental games: the study of strategic interaction,” in *International Series in Experimental Social Psychology*, Vol. XII (Oxford: Pergamon Press), 301. ISBN: 0-08-026069-1
- Fishburn, P. (1974). Lexicographic orders, utilities and decision rules: a survey. *Manage. Sci.* 20, 1442–1471. doi: 10.1287/mnsc.20.11.1442
- Gigerenzer, G., and Todd, P. (1999). *Simple Heuristics that Make us Smart*. New York, NY: Oxford University Press.
- Glimcher, P. (2011). *Foundations of Neuroeconomic Analysis*. New York, NY: Oxford University Press.
- Guillermo, O. (1995). *Game Theory*. San Diego, CA: Academic Press.
- Heckhausen, J., and Heckhausen, H. (2008). *Motivation and Action*. New York, NY: Cambridge University Press. doi: 10.1017/CBO9780511499821
- Johnson, D., Stopka, P., and Bell, J. (2002). Individual variation evades the prisoner’s dilemma. *BMC Evol. Biol.* 2:15. doi: 10.1186/1471-2148-2-15
- Kahneman, D., and Tversky, A. (1979). Prospect theory: an analysis of decision under risk. *Econometrica* 47, 263–292. doi: 10.2307/1914185
- Kaisers, M., and Tuyls, K. (2010). “Frequency adjusted multiagent Q-learning,” in *The Ninth International Conference on Autonomous Agents and Multi-Agent Systems* (Toronto, ON).
- Kaminka, G., Erusalimchik, D., and Kraus, S. (2010). “Adaptive multi-robot coordination: a game-theoretic perspective,” in *IEEE International Conference on Robotics and Automation*. (Anchorage, AK: IEEE).
- Keeney, R. L., and Raiffa, H. (1976). *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*. New York, NY: Wiley.
- Kuhlman, D., and Marshello, A. (1975). Individual differences in game motivation as moderators of preprogrammed strategy effects in prisoner’s dilemma. *J. Pers. Soc. Psychol.* 32, 922–931. doi: 10.1037/0022-3514.32.5.922
- Kuhlman, D., and Wimberley, D. (1976). Expectations of choice behavior held by cooperators, competitors and individualists across four classes of experimental game. *J. Pers. Soc. Psychol.* 34, 69–81. doi: 10.1037/0022-3514.34.1.69
- Li, S., Wang, Z.-J., Rao, L.-L., and Li, Y.-M. (2010). Is there a violation of Savage’s sure-thing principle in the prisoner’s dilemma game. *Adapt. Behav.* 18, 377–385. doi: 10.1177/1059712310366040
- Loyall, A. B. (1997). *Believable agents: building interactive personalities*. Ph.D. thesis. Pittsburgh, PA: Carnegie Mellon University.
- Maher, M. L., Merrick, K., and Saunders, R. (2007). “From passive to proactive design elements: incorporating curious agents into building design,” in *CAADFutures* (Sydney, NSW), 447–460.
- Maynard-Smith, J. (1982). *Evolution and the Theory of Games*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511806292
- McClelland, J., and Watson, R. I. (1973). Power motivation and risk-taking behaviour. *J. Pers.* 41, 121–139.
- McKelvey, R., and Palfrey, T. (1992). An experimental study of the centipede game. *Econometrica* 60, 803–836. doi: 10.2307/2951567
- Meng, Y. (2008). Multi-robot searching using game theoretic based approach. *Adv. Robot. Syst.* 5, 341–350.
- Merrick, K., and Maher, M. L. (2009). *Motivated Reinforcement Learning: Curious Characters for Multiuser Games*. Berlin: Springer. doi: 10.1007/978-3-540-89187-1
- Merrick, K., and Shafi, K. (2011). Achievement, affiliation and they can accurately simulate human decision-making under certain constrained conditions. Specifically, their work focused on single-shot decisions by individual agents. The work in this paper provides a foundation for extending their work to scenarios in which agents interact. In future, such simulations may permit us to examine hypotheses about how individuals with different motives may behave during strategic interactions.
- Key research challenges in this area include understanding the ranges of optimally motivating incentives that best represent motivation types such as power, affiliation and achievement motivated individuals. In practice it seems that there is significant overlap between individuals in the three groups. In addition, motivation psychologists have identified hybrid profiles where more than one motive is dominant (Heckhausen and Heckhausen, 2008). For example in the leadership profile both power and achievement motivation are believed to have approximately equal strength. In terms of the work in this paper, this would mean that agents have more than one optimally motivating incentive. Exploration of profiles such as this is a direction for future work that can provide insight into both the role of motivation in humans and its modeling in artificial systems.

ACKNOWLEDGMENTS

This work was supported by a UNSW@ADFA Early Career Researcher Grant: UNSWA SIR30 Z6300 0000 00 PS23595.

- power: motive profiles for artificial agents. *Adapt. Behav.* 9, 40–62. doi: 10.1177/1059712310395953
- Nash, J. (1950). Equilibrium points in n-person games. *Proc. Natl. Acad. Sci. U.S.A.* 36, 48–49. doi: 10.1073/pnas.36.1.48
- Parsons, S., and Wooldridge, M. (2002). Game theory and decision theory in multi-agent systems. *Auton. Agent. Multi. Agent. Syst.* 5, 243–254. doi: 10.1023/A:1015575522401
- Pita, J., Jain, M., Tambe, M., Ordóñez, F., and Kraus, S. (2010). Robust solutions to Stackelberg games: addressing bounded rationality and limited observations in human cognition. *Artificial Intelligence* 174, 1142–1171. doi: 10.1016/j.artint.2010.07.002
- Poundstone, W. (1992). *Prisoner's Dilemma*. New York, NY: Doubleday.
- Rapoport, A. (1967). Exploiter, leader, hero, martyr. *Behav. Sci.* 12, 81–84. doi: 10.1002/bs.3830120202
- Rapoport, A., and Chammah, A. (1965). *Prisoner's Dilemma*. USA: University of Michigan Press.
- Sandholm, T. W., and Crites, R. H. (1996). Multiagent reinforcement learning in the iterated prisoner's dilemma. *BioSystems* 37, 47–166. doi: 10.1016/0303-2647(95)01551-5
- Simkins, C., Isbell, C., and Marquez, N. (2010). “Deriving behavior from personality: a reinforcement learning approach,” in *International Conference on Cognitive Modelling* (Philadelphia, PA), 229–234.
- Terhune, K. W. (1968). Motives, situation and interpersonal conflict within prisoner's dilemma. *J. Pers. Soc. Psychol., Monogr. Suppl.* 8, 1–24. doi: 10.1037/h0025594
- Valluri, A. (2006). Learning and cooperation in sequential games. *Adapt. Behav.* 14, 195–209. doi: 10.1177/105971230601400304
- Van Knippenberg, B., and Van Knippenberg, D. (2005). Leader self-sacrifice and leadership effectiveness: the moderating role of leader prototypicality. *J. Appl. Psychol.* 90, 25–37. doi: 10.1037/0021-9010.90.1.25
- Van Run, G., and Liebrand, W. (1985). The effects of social motives on behavior in social dilemmas in two cultures. *J. Exp. Soc. Psychol.* 21, 86–102. doi: 10.1016/0022-1031(85)90008-3
- Vassiliades, V., and Christodoulou, C. (2010). “Multiagent reinforcement learning in the iterated prisoner's dilemma: fast cooperation through evolved payoffs,” in *The International Joint Conference on Neural Networks* (Barcelona).
- Von Neumann, J., and Morgenstern, O. (1953). *Theory of Games and Economic Behavior*. Princeton, NJ: Princeton University Press.
- Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.
- Received:* 07 May 2013; *accepted:* 07 October 2013; *published online:* 30 October 2013.
- Citation:* Merrick KE and Shafi K (2013) A game theoretic framework for incentive-based models of intrinsic motivation in artificial systems. *Front. Psychol.* 4:791. doi: 10.3389/fpsyg.2013.00791
- This article was submitted to Cognitive Science, a section of the journal Frontiers in Psychology.*
- Copyright © 2013 Merrick and Shafi. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.*



Imitation learning based on an intrinsic motivation mechanism for efficient coding

Jochen Triesch *

Department of Neuroscience, Frankfurt Institute for Advanced Studies, Frankfurt, Germany

Edited by:

Gianluca Baldassarre, Italian National Research Council, Italy

Reviewed by:

Minoru Asada, Osaka University, Japan

Pierre-Yves Oudeyer, Institut National de Recherche en Informatique et en Automatique, France

Tjeerd C. Andringa, University of Groningen, Netherlands

***Correspondence:**

Jochen Triesch, Department of Neuroscience, Frankfurt Institute for Advanced Studies,

Ruth-Moufang-Str. 1, 60438 Frankfurt, Germany

e-mail: triesch@fias.uni-frankfurt.de

A hypothesis regarding the development of imitation learning is presented that is rooted in intrinsic motivations. It is derived from a recently proposed form of intrinsically motivated learning (IML) for efficient coding in active perception, wherein an agent learns to perform actions with its sense organs to facilitate efficient encoding of the sensory data. To this end, actions of the sense organs that improve the encoding of the sensory data trigger an internally generated reinforcement signal. Here it is argued that the same IML mechanism might also support the development of imitation when general actions beyond those of the sense organs are considered: The learner first observes a tutor performing a behavior and learns a model of the behavior's sensory consequences. The learner then acts itself and receives an internally generated reinforcement signal reflecting how well the sensory consequences of its own behavior are encoded by the sensory model. Actions that are more similar to those of the tutor will lead to sensory signals that are easier to encode and produce a higher reinforcement signal. Through this, the learner's behavior is progressively tuned to make the sensory consequences of its actions match the learned sensory model. I discuss this mechanism in the context of human language acquisition and bird song learning where similar ideas have been proposed. The suggested mechanism also offers an account for the development of mirror neurons and makes a number of predictions. Overall, it establishes a connection between principles of efficient coding, intrinsic motivations and imitation.

Keywords: intrinsic motivation, imitation, efficient coding, active perception, language development, bird song, mirror neuron, perceptual fluency

1. INTRODUCTION

Imitation is a powerful form of learning where an agent acquires a skill from observing the skill being performed by a second agent. This can dramatically speed up the learning of useful behaviors compared to random exploration (Miller and Dollard, 1941). In the animal learning literature, imitation has been defined as “the copying of a novel or otherwise improbable act or utterance, or some act for which there is clearly no instinctive tendency” (Thorpe, 1963), but many other more or less stringent definitions exist. Many authors reserve the term imitation to situations where the behavior in question is not yet in the behavioral repertoire of the imitating agent (Clayton, 1978), but assessing the behavioral repertoire of an animal is in itself problematic. In the following, I will simply use imitation as an umbrella term for various forms of social learning and highlight important distinctions in the context of specific examples.

Despite many years of research, the origin and development of imitation abilities in animals and humans are still poorly understood (Heyes, 2001). While some theories have proposed that the ability to imitate relies on sophisticated innate mechanisms (Meltzoff and Moore, 1997), other accounts have emphasized the role of generic learning mechanisms for the development of imitative behaviors (Miller and Dollard, 1941; Gewirtz, 1969). Recent learning accounts considering possible underlying neurobiological mechanisms have rested on associative (Hebbian)

learning (Heyes and Ray, 2000; Keysers and Perrett, 2004) or reinforcement learning (Triesch et al., 2007). These are sufficient for the development of a simple form of imitation also called *response facilitation*, where the agent learns to map the observation of a behavior performed by a second agent onto an already existing motor representation for performing the same behavior. This motor representation could already be present at birth or have been learned previously through random exploration of movement possibilities, often referred to as *babbling*. Importantly, however, these accounts have difficulties explaining the development of what is sometimes called *true imitation*, where the to-be-learned skill is not yet in the behavioral repertoire of the developing agent. This is the much more difficult and interesting case, because it addresses how imitation could accelerate the acquisition of novel skills.

An important example is speech acquisition, where the infant learns to produce utterances from her native language based on interactions with her caregivers. Infants are capable of statistical learning and readily discover statistical patterns of their native language, but also the social interaction with caregivers is critical for normal development of speech, see Kuhl (2004) for review. A closely related case is the acquisition of songs in certain species of song birds. This learning has been related to human language learning (Marler, 1970; Doupe and Kuhl, 1999) and is used as a model system for it. As early as 1773 it was shown that birds learn

their song(s) from experience during development (Barrington, 1773). For example, male juvenile zebra finches usually learn to sing a song that closely resembles that of their father. The learning proceeds in two phases. During a first phase of purely sensory learning, the juvenile bird is suspected to form an auditory template of the father's (or other social tutor's) song (Baptista and Petrinovich, 1984; Konishi, 2010). During a second phase of sensory-motor learning, the bird learns to produce a song to match the learned template. Depending on the species, the sensory and sensory-motor phases may or may not overlap. Presently it is still unclear through what precise mechanisms the juvenile bird manages to better and better approximate the father's song. Here I discuss how a recently proposed intrinsically motivated learning (IML) mechanism for efficient coding in active perception might be generalized for this form of imitation learning. This suggests that principles of efficient sensory coding may be a foundation for song learning in birds and speech acquisition in humans.

Intrinsic motivations have recently come into focus as important driving forces in the development of complex behaviors (Baldassarre and Mirolli, 2013). While there is still much debate about the correct definition of intrinsic motivations (Baldassarre, 2011), the term is usually used when referring to behaviors such as play or other "curious" exploration of the environment that seem unrelated to any immediate "extrinsic" goal such as the acquisition of food. This *hypothesis* article does not propose any specific computational model nor does it present any empirical results. It is merely discussing the new hypothesis in the context of existing work. In the following, I briefly review a recently proposed form of IML for efficient sensory coding in active perception. Then I show how a generalization of this mechanism may account for the development of imitative behaviors. This also suggests a mechanism for the development of mirror neurons. Finally, I discuss predictions that the proposed mechanism makes.

2. INTRINSICALLY MOTIVATED LEARNING FOR EFFICIENT CODING IN ACTIVE PERCEPTION

The efficient coding hypothesis posits that sensory systems strive to encode sensory information in an efficient manner by exploiting the statistical structure and redundancies present in the sensory data (Attneave, 1954; Barlow, 1961). Since its first formulation, numerous aspects of sensory coding have been successfully explained in this context. This includes research on how early visual representations can be understood as adaptations to the statistics of natural images (Simoncelli and Olshausen, 2001) as well as related findings in the auditory (Smith and Lewicki, 2006) and olfactory (Perez-Orive et al., 2002) modalities. While this research program has been highly successful, it has typically neglected the active nature of perception. In particular, the statistics of sensory signals are a result of both the natural environment and the organism's behavior. This implies that the behavior of the organism and in particular the movement of the sense organs could be utilized to make the encoding of sensory information more efficient.

Along these lines and inspired by previous work from Schmidhuber (2009) proposing compression progress as an objective for IML, Zhao et al. (2012) have recently presented a

model that learns to efficiently encode visual input from two eyes, see **Figure 1A**. Their approach proposes a form of IML using an internally generated reinforcement signal for learning efficient coding strategies in active perception. The method works as follows: A sensory model learns to encode sensory data, while a reinforcement learner generates actions of the sense organs that help the agent to encode the sensory data efficiently. To this end, an internally generated reinforcement signal is given to the reinforcement learner that reflects how well the sensory model is able to encode the input.

In the context of binocular vision Zhao et al. (2012) have shown that this mechanism elegantly explains the joint development of an efficient representation for stereo disparity in the sensory model and an accurate controller for vergence eye movements. In this setting, the system discovers that it is useful (intrinsically rewarding) to verge both eyes onto a common physical point, because then the sensory model is able to encode the data more efficiently. This is because the images from both eyes become more redundant and their joint encoding by the sensory model becomes more accurate. We may think of this in terms of the affordance concept. The observation of a certain disparity at the center of gaze is found to *afford* a certain vergence command that will lead to an improved representation of this input.

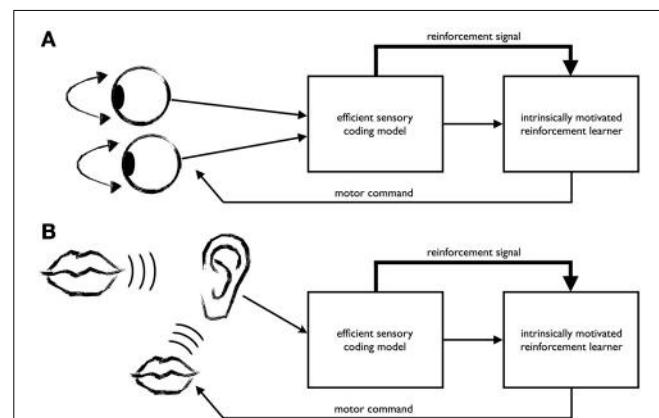


FIGURE 1 | The recently proposed intrinsically motivated learning architecture for efficient coding in active perception (A) also gives rise to the development of imitation (B). (A) The learning architecture comprises an efficient coding model for the sensory input and an intrinsically motivated reinforcement learning mechanism for generating behavior. In the example of Zhao et al. (2012), the efficient coding model learns a sparse code for binocular images, while the reinforcement learner generates vergence eye movements. To this end, it receives from the sensory coding model a representation of the sensory input (thin arrow) and an internally generated reward signal reflecting how well the sensory model could encode the binocular input (thick arrow). Both the sensory coding model and the reinforcement learner try to optimize the encoding of the data. The system discovers that the input data can be encoded most efficiently when vergence commands are used to minimize binocular disparity. (B) The learner acquires an efficient encoding of speech signals provided by a tutor (big mouth). When the learner starts babbling (small mouth), the resulting acoustic signals are encoded by the sensory model that has been tuned to the tutor's speech. Signals that are easy to encode for the sensory model because the utterance sounds similar to the tutor's speech will produce a high reinforcement signal. Through this, the system's utterances are progressively driven to approximate the tutor's speech.

Importantly, the learning of the sensory model and the eye movement control develop jointly in this approach, driven by the identical objective of encoding the data efficiently. This mechanism has been shown to lead to fully autonomous and self-calibrating development of binocular vision and has been validated on a real robot (Lonini et al., 2013). More recently, it has also been extended to the development of smooth pursuit eye movements. Whether this approach can be extended to actions beyond eye movements is still an open question.

The central assumption of this approach is the existence of an internally generated reinforcement signal that encourages movements of the sense organs leading to an improved encoding of the sensory stimulus. Research on perceptual fluency supports the plausibility of this assumption. It has been found that the ease of processing of a sensory stimulus is related to positive affect (Reber et al., 1998). Assuming that the ease of processing reflects the quality of encoding of the stimulus by the sensory model, then easy to encode stimuli should produce positive affect. This positive affect may be due to the proposed internally generated reinforcement signal.

One point requires some discussion, however. Simply trying to behave such that the incoming sensory signals are encoded most easily might drive the agent to more or less abolish sensory input. In the case of visual perception, the agent could simply close the eyes or stare at a blank wall. This would make the sensory signals be encoded most easily, but is of little use otherwise. There are several ways to avoid this. A first solution is to introduce a separate mechanism for selecting *what* the agent will look at, while the described IML mechanism ensures that *how* the target object is being looked at is most efficient. For example, an attention mechanism selects what object in the scene should be looked at, while the proposed IML mechanism ensures that this particular object is well represented through vergence, smooth pursuit, and possibly other eye, head, and body movements. At the same time, it provides an optimized sensory encoding of the stimulus by properly taking into account the statistics of the sensory signals resulting from these movements. A second solution to the problem is to measure the ease of encoding of the sensory data in relation to some notion of the complexity of the data or the amount of information it contains. For example, the sensory signals resulting from staring at the blank wall may indeed be easy to encode (e.g., lead to a low reconstruction error of a generative model), but they may contain very little information to start with. There are various ways of making these notions mathematically precise, but the details are not important for the present paper.

Having introduced the recently proposed IML mechanism for efficient coding in active perception, we are now ready to consider its connection to imitation learning, which will require us to generalize it from movements of the sense organs to other motor acts.

3. HOW INTRINSICALLY MOTIVATED LEARNING FOR EFFICIENT CODING MAY SUPPORT IMITATION

The mechanism for IML in active perception discussed above could also lead to the development of a form of imitation learning, as illustrated in **Figure 1B**. Consider the example of an infant

faced with the problem of acquiring speech by imitating the utterances of her caregivers (or that of a juvenile song bird learning the father's song). Let's assume that at a certain point in development the infant has already learned a reasonably good sensory representation of what her native language sounds like Kuhl (2004). This representation will continue to improve with age and experience. When the infant vocalizes, her utterances will be processed by her own auditory system, which has already been tuned toward the sounds and words of her mother tongue. According to the IML mechanism described above, utterances that sound more like her mother tongue will be more easily encoded by her auditory system, which will lead to the generation of a higher reinforcement signal compared to utterances that sound dissimilar from her mother tongue. Thus, over time, the infant will adapt her utterances to the language she is exposed to driven by her intrinsic motivation to behave in such a way that the sensory data are encoded easily for her auditory system. Importantly, this suggests that language specific information could enter the babbling process early on, with each utterance being evaluated in the light of already learned sensory representations. We will return to this point in the Discussion.

An important question in this context is how the sensory model will learn to encode the caregiver's speech and when exactly the infant's speech will be easy to encode for the sensory model. The caregiver's utterances will necessarily sound different from the infant's utterances due to the different structure of their vocal tracts. For example, it is not clear why the sound of a certain vowel produced by the infant with her vocal tract should be easy to encode for her auditory system, if this has been tuned to speech of her caregiver, whose vowels will generally differ in fundamental frequency and other parameters. For the case of vowel acquisition in the context of infant caregiver interactions, it has been argued that an *automirroring bias* can overcome this difficulty Ishihara et al. (2009); Miura et al. (2012).

3.1. RELATIVE TIMING OF SENSORY AND MOTOR LEARNING

For the proposed IML mechanism it may be maladaptive for the learner to produce utterances at an excessive rate right after birth. If a sensory representation properly reflecting the correct target language (or song) is not acquired first, then the learner's auditory representation may become tuned to or even dominated by its own utterances. According to the proposed IML mechanism, the learner would then find rewarding whatever it is producing itself. This could potentially slow down learning of the native language. Enforcing a sufficient amount of passive exposure to the language may avoid this problem.

Similarly, reducing plasticity in sensory areas at the end of a critical period and before the onset of vocalizations may also alleviate this problem, because it prevents the sensory representation from becoming dominated by the sensory consequences of the agent's own actions.

An alternative solution to the problem would be to reduce or switch off sensory learning during one's own vocalizations. Instead, the auditory feedback could be used to train a forward model that predicts the auditory feedback based on an efference copy of the motor signals. Note that an accurate forward model allows planning and off-line learning without the need for

producing actual motor output and observing the consequences. This can dramatically speed up learning (Sutton and Barto, 1998) and could even happen during sleep.

3.2. LEARNING ONE THING OR MANY?

As discussed above, the absence of sensory input might be particularly easy to encode for the sensory system. This might lead the infant to not vocalize at all. Several solutions are conceivable. First, as suggested above the quality of encoding of the sensory model could be relative to the complexity of the sensory input or the amount of information contained in it. In this way, the situation of not babbling at all could be made comparatively undesirable. Second, a mechanism reinforcing the learning of novel cause and effect relationships or the “discovery of novel actions” (Redgrave and Gurney, 2006) could foster varied babbling. Third and maybe most obviously, the infant may want to communicate.

The question remains what and how many different things might be acquired through this IML mechanism. Note that while some bird species only learn a single song that “crystallizes” during development, others learn thousands of utterances during their life time (Catchpole and Slater, 2003) as do humans. If the sensory model allowed for only a single song “template” to be stored, this might explain why only a single song is learned. If, however, the sensory model had a high capacity for storing many acoustic patterns with high fidelity, then a large repertoire of actions would be learned with this mechanism. In general, for any kind of sensory model there will be a trade-off: given a fixed storage capacity more patterns can only be stored at the cost of storing them with smaller fidelity. Such differences could contribute to the varied vocabulary sizes in different species of song birds.

3.3. CONTEXT DEPENDENCE

The mechanism described thus far will allow an agent to learn to imitate a range of utterances or behaviors whose sensory consequences match those of its learned sensory model. In the simplest case, however, all of these behaviors will appear equally “good” in any situation, i.e., what vocalization is performed would not necessarily depend on the current context. This could lead to behaviors being produced in inappropriate contexts. How could the agent learn to generate a certain behavior only in the appropriate context?

One solution is certainly through instrumental learning. If, say, the behavior has undesirable consequences in the present context, its execution may be made less probable because of this. A second solution to the problem is that during learning of the sensory model, contextual information is also integrated into the representation. Thus, the model will not be a purely sensory model anymore but a sensory-plus-context model. Specifically, if during the sensory-only phase of development, the infant or the song bird hears an utterance only in a specific context, then the developing sensory-plus-context model may encode this relationship. Thereby, if the learner generates the behavior in the same context, this will be particularly easy to encode for the sensory-plus-context model. Conversely, if the behavior is produced in a different context, this will be less easy to encode for the

sensory-plus-context model, because there is a mismatch between the context and the sensory input. Obviously, relevant contexts are also perceived based on sensory, e.g., visual information. Thus a strict separation of sensory information and context may not always be possible. Interestingly, the context could be the presence of a certain object to which the infant pays attention. In this case, an initial association between the visual appearance of the object, its acoustic label, and the motor representation for generating the acoustic label can be established. In this situation, the presence of the object would *afford* producing the object’s name.

4. DEVELOPMENT OF MIRROR NEURONS

Mirror neurons are a class of neurons first observed in the pre-motor cortex of monkeys (Gallese et al., 1996) whose defining characteristic is that they can be activated if the monkey observes another agent performing a certain behavior or if the monkey plans and executes the same behavior. Because of this, they have been implicated in action understanding, imitation, empathy and language acquisition (Rizzolatti and Arbib, 1998; Gallese et al., 2004; Rizzolatti and Craighero, 2004). While originally discovered in monkeys, there is converging evidence for a mirror neuron system in humans (Iacoboni et al., 1999) and song birds (Prather et al., 2008). While the question how mirror neurons could support imitation has received much interest (Iacoboni et al., 1999; Iacoboni, 2005, 2009), comparatively little work has investigated how mirror neurons develop ontogenetically and what learning processes drive this development (Heyes, 2010).

Complementary mechanisms have been proposed for the development of mirror neurons based on generic learning principles. The most popular one is that mirror neurons develop through associative learning mechanisms such as Hebbian learning (Heyes and Ray, 2000; Keysers and Perrett, 2004; Heyes et al., 2005; Catmur et al., 2007; Cooper et al., 2013). A second mechanism is that mirror neurons could develop through reward-dependent (instrumental, reinforcement) learning (Triesch et al., 2007). We will take a look at both mechanisms before describing a new one based on IML for efficient coding, which combines aspects of the other two.

4.1. HEBBIAN DEVELOPMENT OF MIRROR NEURONS

Hebbian accounts works as follows (Heyes and Ray, 2000; Keysers and Perrett, 2004; Del Giudice et al., 2009). In the case of behaviors whose sensory consequences are easily observed such as seeing one’s own reaching movement or hearing one’s own utterances, it is assumed that Hebbian learning forms associations between simultaneously active sensory and motor representations for already learned skills. As a result, neurons involved in the execution of a specific behavior receive strong excitatory connections from neurons representing its sensory consequences and vice versa. When another agent is then observed performing the same action, the same sensory representations will be triggered due to their ability to generalize to similar sensory stimuli. It has been argued that such generalization ability may stem from maturational constraints of the visual system Nagai et al. (2011). The activated sensory representation then excites the corresponding motor representation via the associative connections learned

through the Hebbian mechanism. Through this the motor representation has obtained mirror properties: it is activated by planning or executing a behavior and by merely observing it in another agent.

The situation is more difficult for behaviors where the agent cannot fully perceive the sensory consequences of its actions as in the generation of facial expressions. For such “opaque” cases it is assumed that the agent learns to imitate by first being imitated by another agent—usually the caregiver. For example, when an infant smiles and his mother imitates the smile, the infant can learn to associate the visual representation of the mother’s smiling face with her own motor representation for smiling. Again, the motor representation assumes mirror properties due to Hebbian learning. While overall the account appears plausible, a limitation is that it only develops mirror representations for skills that have already been learned. The learning of novel behaviors is left to random exploration which is very inefficient when many motor degrees-of-freedom are involved as is the case in speech or song production, i.e., when learning takes place in a high-dimensional space.

4.2. REWARD-DRIVEN DEVELOPMENT OF MIRROR NEURONS

In the reward-based learning account, the agent discovers that performing a certain behavior is useful whenever it sees another agent perform this behavior. For example, when a developing monkey observes a conspecific grasping a peanut from a source, the resulting sensory representation can become associated with the monkey’s own motor plan for grasping a peanut from the same source, which is inherently rewarding—especially when hungry. Note that this mechanism does not require the ability to observe the sensory appearance of one’s own action, but only whether it leads to a positive, i.e., reinforcing outcome. Circumstantial evidence for the importance of reward-driven learning in the development of mirror neurons comes from a recent finding that mirror neurons in monkey premotor area F5 are modulated by the value the monkey assigns to a grasped object (Caggiano et al., 2012).

The reward-driven account was studied in greatest detail in the context of gaze following, where an agent learns to look where others are looking. This is an example of a behavior where the sensory appearance of the behavior cannot be observed while the agent performs it. Triesch et al. (2007) proposed a computational model for the development of gaze following and showed that it produced mirror neurons for looking behaviors. It also explained various other aspects of the development of gaze following (Jasso et al., 2012). The existence of mirror neurons was the central prediction of the model and it was later confirmed neurophysiologically (Shepherd et al., 2009).

Interestingly, the reward-driven learning mechanism also predicts the possibility of generalized mirror neurons (Triesch et al., 2007). An agent may discover that it is useful to perform some action A whenever another agent is observed performing an action B. Gaze following represents a simple example of this: when two agents face each other, proper gaze following requires the learning agent to turn the head to his left if the model is observed turning the head to its right. Thus, not the physical appearance of the movement matters, but the goal of the

action: where should I look? Through the reward driven learning mechanism an association can be learned from the sensory representation corresponding to the observation of the other agent performing action B and one’s own motor representation of action A. This would lead to generalized mirror neurons for which the observed action triggering them is not necessarily identical to the action being generated.

4.3. INTRINSICALLY MOTIVATED DEVELOPMENT OF MIRROR NEURONS

The proposed IML mechanism integrates ideas from the Hebbian and the reward-based accounts. Like the Hebbian mechanism, it requires that the sensory consequences of the actions can be perceived. The development of mirror neurons could proceed along the following steps. (1) During sensory-only learning, a sensory model of various behaviors produced by the tutor is learned. Associated with this model, we assume that there will be populations of neurons specific to the perception of these different behaviors. (2) During the sensory-motor phase, the learner acquires motor representations that produce the same sensory consequences by virtue of the proposed IML mechanism. This involves the learner’s reward system, but the reinforcement signals are internally generated. In the end, specific motor representations and the associated populations of neurons will code for specific behaviors. (3) Since these motor representations trigger specific sensory consequences, Hebbian learning mechanisms can establish a bidirectional association between the motor representation and the sensory representation. Through this, the sensory representation will acquire some motor properties and the motor representation will acquire some sensory properties. The clear distinction between sensory and motor representations dissolves and neurons with mirror properties develop: They are active when their sensory representation is triggered during observation of the behavior of another agent and during planning and execution of the corresponding behavior. Note that, the three steps could also overlap in time.

The computational benefit of the IML mechanism over the Hebbian mechanism is that the discovery of new skills is not left to random exploration, but occurs under guidance from the sensory model. Exploration is focused on those behaviors that produce similar sensory consequences as the behavior of conspecifics. The computational advantage over the reward-based mechanism is similar. The discovery of new skills does not require an external reward such as the peanut in the above example, but guarantees that matching one’s behavior with that of a conspecific is intrinsically rewarding. This seems to better reflect the true nature of at least human imitation.

5. DISCUSSION

I have described how a recently proposed mechanism for IML for efficient coding in active perception can be generalized to support imitative learning. In addition, a corresponding account for the development of mirror neurons was presented. It combines previous proposals based on associative Hebbian learning and instrumental or reinforcement learning in the framework of IML. These mechanisms represent parallel pathways through which mirror neurons can be acquired. Once established through

either of these mechanisms, it is easy to see how mirror neurons could contribute to various forms of imitation including automatic imitation (Heyes, 2010) and vocal mimicry.

The IML mechanism proposed here is compatible with many previous theoretical accounts and computational models of song bird learning. A full review of these works is beyond the scope of this article. Existing works typically assume that a reinforcement signal is derived from matching auditory feedback to a stored sensory template (Doya and Sejnowski, 1995; Troyer and Doupe, 2000). Here I have proposed that such a reward signal could be derived from an evaluation of how well the auditory feedback is encoded by a sensory model. This distinction is admittedly subtle, but it connects the present approach to theories on efficient coding and sparse coding models as we have used in our work on the role of the same IML mechanism in active perception (Zhao et al., 2012; Lonini et al., 2013). This may be important, since neural representations in certain parts of the song system are known to be very sparse (Hahnloser et al., 2002).

The examples of human language acquisition and bird song learning are special in that the sensory consequences of the behavior are readily perceived. Obviously, the proposed mechanism can be extended to other actions that are easily perceived such as manual actions. For other actions such as facial expressions, this is not straight forward (unless a mirror is available). Learning to imitate facial expressions may require other mechanisms such as being imitated by caregivers (Heyes, 2001) or rely on reinforcement learning mechanisms and social feedback.

The presented mechanism is rooted in the efficient coding hypothesis. As such, it somewhat downplays the importance of social feedback during speech and song acquisition. But the social context in which learning takes place is known to play a very important role both in human language acquisition and bird song learning Goldstein et al. (2003); Kuhl et al. (2003). In the words of Goldstein and Schwade (2008): “infants’ prelinguistic vocalizations, and caregivers’ reactions to those immature sounds, create opportunities for social learning that afford infants knowledge of phonology.”

The proposed IML mechanism also shares some aspects of previous work on imitation in the developmental robotics literature. For instance, (Gaussier et al., 1998) and (Andry et al., 2001) propose a robot where a mechanism of “cognitive homeostasis” would give rise to imitative behaviors. Due to a “perceptual ambiguity” the robot may mistake an optic flow field caused by observing a moving agent with the flow field produced by its own locomotion. The homeostasis drive would try to minimize the mismatch between the sensory input stream and the robot’s motor commands such that the robot will start moving. This is suggested to lead to an immediate following behavior. They then present experiments with a real robot that has a different prewired following mechanism. It learns to store extended sequences of movements resulting from following another robot or a human if these sequences lead to a reward. In our case, imitation does not emerge from a drive to reduce the mismatch between sensory percepts and own motor commands or from a prewired following mechanism but from a reinforcement signal that favors movements whose sensory consequences can be encoded efficiently by the sensory system.

Kaplan and Oudeyer (2007) have considered an intrinsic motivation for maximizing learning progress and discussed its potential role in the development of imitation. After illustrating how an intrinsic motivation for learning progress allows an agent to tackle progressively more difficult learning problems by discovering “progress niches,” they speculate that such an intrinsic motivation may also contribute to the development of imitation. Specifically, they argue that “(1) the meaningful distinctions necessary for the development of imitation (self, others and objects in the environment) may be the result of discriminations constructed during a progress-driven process and that (2) imitative behavior can more generally be understood as a way of producing actions in order to experience learning progress.” They speculate that at different stages of development infants may engage in different kinds of imitative behaviors because they maximize the infant’s current learning progress. Here we argue that imitative behaviors are reinforced because their sensory consequences can be encoded efficiently by the learner’s sensory model.

How could the proposed IML mechanism be tested experimentally? In the context of human language learning, it suggests that the babbling process might already reflect some aspects of the statistical properties of the language to which the infant has been exposed. This in turn predicts that the babbling process of infants could be shaped by carefully controlling their language input. For example, we may speculate that when caregivers intuitively reply to babbling attempts by uttering “close” words from the target language, they will affect the infant’s sensory model in such a way that the correct pronunciation of the “close” word is reinforced during future babbling attempts. In contrast, replying to infant’s babbling attempts with arbitrary different-sounding words will not produce this effect. Other aspects of child-directed speech such as hyperarticulation are also thought to aid the infant in learning a sensory model of the target language (Kuhl et al., 1997). More research is needed to investigate if and how infants’ babbling is shaped by their developing sensory model of the target language through internally generated reinforcement signals.

In the context of bird song learning, the IML mechanism could be tested most directly by recording from reward circuits in the song bird brain as the animal is learning its song. The most obvious and direct prediction is that utterances sounding more similar to the father’s song will generate a higher reward signal because they are easier to encode for the bird’s auditory system, while utterances sounding dissimilar from the father’s song will generate a lower reward signal because they are harder to encode. By manipulating the auditory feedback the bird is receiving, the causal role of this sensory feedback in learning can be tested. Note, however, that disentangling whether a stronger reinforcement signal is due to an easier encoding of the sensory signals or a greater similarity of the auditory feedback to a stored template may be difficult. To this end, it may be important to consider song bird species learning many different songs.

Next to testing the proposed mechanism and its possible neural implementation in biological experiments, it will also be interesting to apply the idea in the context of robots. For example, future work could try to exploit the proposed IML mechanism for language learning in robots. This will help to identify possible limitations or inconsistencies of the approach. The experiences

gained would help to further develop and refine the current proposal. In conclusion, it is intriguing that the venerable principle of efficient sensory coding may play a central role in sophisticated cognitive phenomena such as imitation and language acquisition.

FUNDING

This work was supported by the Quandt foundation, the European Communities Seventh Framework Programme FP7/2007-2013, Challenge 2 - Cognitive Systems, Interaction, Robotics, under grant agreement No FP7-ICT-IP-231722, project IM-CLeVeR Intrinsically Motivated Cumulative Learning Versatile Robots, the DAAD through the Germany/Hong Kong Joint Research Scheme (project number G HK25/10), and the BMBF through Project Bernstein Fokus: Neurotechnologie Frankfurt, FKZ 01GQ0840.

ACKNOWLEDGMENTS

The author thanks M. Murakami and three anonymous reviewers for helpful comments on previous versions of the manuscript and G. Deák, C. Rothkopf, B. Shi and the partners of the IM-CLeVeR project for their (intrinsic?) motivation to collaborate on related topics.

REFERENCES

- Andry, P., Gaussier, P., Moga, S., Banquet, J.-P., and Nadel, J. (2001). Learning and communication via imitation: an autonomous robot perspective. *Syst. Man Cybern. A* 31, 431–442. doi: 10.1109/3468.952717
- Attneave, F. (1954). Some informational aspects of visual perception. *Psychol. Rev.* 61, 183. doi: 10.1037/h0054663
- Baldassarre, G. (2011). “What are intrinsic motivations? A biological perspective,” in *Proceeding IEEE International Conference on Development and Learning (ICDL)*, (Frankfurt), 1–8. doi: 10.1109/DEVLRN.2011.6037367
- Baldassarre, G., and Mirolli, M., (eds) (2013). *Intrinsically Motivated Learning in Natural and Artificial Systems*. Berlin: Springer. doi: 10.1007/978-3-642-32375-1
- Baptista, L., and Petrinovich, L. (1984). Social interaction, sensitive phases, and the song template hypothesis in the white-crowned sparrow. *Anim. Behav.* 32, 172–181. doi: 10.1016/S0003-3472(84)80335-8
- Barlow, H. B. (1961). Possible principles underlying the transformations of sensory messages. Chapter 13”, in *Sensory Communication*, ed W. Rosenblith (M.I.T. Press), 217–234.
- Barrington, D. (1773). Experiments and observations on the singing of birds. *Philos. Trans. R. Soc. Lond.* 63, 249–291. doi: 10.1098/rstl.1773.0031
- Caggiano, V., Fogassi, L., Rizzolatti, G., Casile, A., Giese, M. A., and Thier, P. (2012). Mirror neurons encode the subjective value of an observed action. *Proc. Natl. Acad. Sci. U.S.A.* 109, 11848–11853. doi: 10.1073/pnas.1205553109
- Catchpole, C. K., and Slater, P. J. (2003). *Bird Song: Biological Themes and Variations*. Cambridge: Cambridge University Press.
- Catmur, C., Walsh, V., and Heyes, C. (2007). Sensorimotor learning configures the human mirror system. *Curr. Biol.* 17, 1527–1531. doi: 10.1016/j.cub.2007.08.006
- Clayton, D. (1978). Socially facilitated behavior. *Q. Rev. Biol.* 53, 373–391. doi: 10.1086/410789
- Cooper, R. P., Cook, R., Dickinson, A., and Heyes, C. M. (2013). Associative (not Hebbian) learning and the mirror neuron system. *Neurosci. Lett.* 540, 28–36. doi: 10.1016/j.neulet.2012.10.002
- Del Giudice, M., Manera, V., and Keysers, C. (2009). Programmed to learn? The ontogeny of mirror neurons. *Dev. Sci.* 12, 350–363. doi: 10.1111/j.1467-7687.2008.00783.x
- Doupe, A., and Kuhl, P. (1999). Birdsong and human speech: common themes and mechanisms. *Annu. Rev. Neurosci.* 22, 567–631. doi: 10.1146/annurev.neuro.22.1.567
- Doya, K., and Sejnowski, T. (1995). A novel reinforcement model of birdsong vocalization learning. *Adv. Neural Inf. Process. Syst.* 7, 101–108.
- Gallese, V., Fadiga, L., Fogassi, L., and Rizzolatti, G. (1996). Action recognition in the premotor cortex. *Brain* 119, 593–609. doi: 10.1093/brain/119.2.593
- Gallese, V., Keysers, C., and Rizzolatti, G. (2004). A unifying view of the basis of social cognition. *Trends Cogn. Sci.* 8, 396–403. doi: 10.1016/j.tics.2004.07.002
- Gaussier, P., Moga, S., Quoy, M., and Banquet, J.-P. (1998). From perception-action loops to imitation processes: a bottom-up approach of learning by imitation. *Appl. Artif. Intell.* 12, 701–727. doi: 10.1080/088395198117596
- Gewirtz, J. L. (1969). Mechanisms of social learning: Some roles of stimulation and behavior in early human development. *Handbook Soc. Theory Res.* 57–212.
- Goldstein, M. H., King, A. P., and West, M. J. (2003). Social interaction shapes babbling: testing parallels between birdsong and speech. *Proc. Natl. Acad. Sci. U.S.A.* 100, 8030–8035. doi: 10.1073/pnas.1332441100
- Goldstein, M. H., and Schwade, J. A. (2008). Social feedback to infants' babbling facilitates rapid phonological learning. *Psychol. Sci.* 19, 515–523. doi: 10.1111/j.1467-9280.2008.02117.x
- Hahnloser, R. H., Kozhevnikov, A. A., and Fee, M. S. (2002). An ultra-sparse code underlies the generation of neural sequences in a songbird. *Nature* 419, 65–70. doi: 10.1038/nature00974
- Heyes, C. (2001). Causes and consequences of imitation. *Trends Cogn. Sci.* 5, 253–261. doi: 10.1016/S1364-6613(00)01661-2
- Heyes, C. (2010). Where do mirror neurons come from? *Neurosci. Biobehav. Rev.* 34, 575–583. doi: 10.1016/j.neubiorev.2009.11.007
- Heyes, C., Bird, G., Johnson, H., and Haggard, P. (2005). Experience modulates automatic imitation. *Cogn. Brain Res.* 22, 233–240. doi: 10.1016/j.cogbrainres.2004.09.009
- Heyes, C. M., and Ray, E. D. (2000). What is the significance of imitation in animals? *Adv. Study Behav.* 29, 215–245. doi: 10.1016/S0065-3454(08)60106-0
- Iacoboni, M. (2005). Neural mechanisms of imitation. *Curr. Opin. Neurobiol.* 15, 632–637. doi: 10.1016/j.conb.2005.10.010
- Iacoboni, M. (2009). Neurobiology of imitation. *Curr. Opin. Neurobiol.* 19, 661–665. doi: 10.1016/j.conb.2009.09.008
- Iacoboni, M., Woods, R. P., Brass, M., Bekkering, H., Mazziotta, J. C., and Rizzolatti, G. (1999). Cortical mechanisms of human imitation. *Science* 286, 2526–2528. doi: 10.1126/science.286.5449.2526
- Ishihara, H., Yoshikawa, Y., Miura, K., and Asada, M. (2009). How caregiver's anticipation shapes infant's vowel through mutual imitation. *Auton. Mental Dev. IEEE Trans.* 1, 217–225. doi: 10.1109/TAMD.2009.2038988
- Jasso, H., Triesch, J., Deák, G., and Lewis, J. (2012). A unified account of gaze following. *IEEE Trans. Auton. Mental Dev.* 4, 257–272. doi: 10.1109/TAMD.2012.2208640
- Kaplan, F., and Oudeyer, P.-Y. (2007). “The progress-drive hypothesis: an interpretation of early imitation,” in *Models and Mechanisms of Imitation and Social Learning: Behavioural, Social and Communication Dimensions*, eds C. Nehaniv and K. Dautenhahn (Cambridge: Cambridge University Press), 361–377.
- Keysers, C., and Perrett, D. (2004). Demystifying social cognition: a Hebbian perspective. *Trends Cogn. Sci.* 8, 501–507. doi: 10.1016/j.tics.2004.09.005
- Konishi, M. (2010). From central pattern generator to sensory template in the evolution of birdsong. *Brain Lang.* 15, 18–20. doi: 10.1016/j.bandl.2010.05.001
- Kuhl, P. K. (2004). Early language acquisition: cracking the speech code. *Nat. Rev. Neurosci.* 5, 831–843. doi: 10.1038/nrn1533
- Kuhl, P. K., Andruski, J. E., Chistovich, I. A., Chistovich, L. A., Kozhevnikova, E. V., et al. (1997). Cross-language analysis of phonetic units in language addressed to infants. *Science* 277, 684–686. doi: 10.1126/science.277.5326.684
- Kuhl, P. K., Tsao, F.-M., and Liu, H.-M. (2003). Foreign-language experience in infancy: effects of short-term exposure and social interaction on phonetic learning. *Proc. Natl. Acad. Sci. U.S.A.* 100, 9096–9101. doi: 10.1073/pnas.1532872100
- Lonini, L., Zhao, Y., Chandrashekhariah, P., Shi, B. E., and Triesch, J. (2013). “Autonomous learning of active multi-scale binocular vision,” in *Development and Learning (ICDL), 2013 IEEE International Conference on*, (Osaka), 1–6.
- Marler, P. (1970). Birdsong and speech development: Could there be parallels? There may be basic rules governing vocal learning to which many species conform, including man. *Am. Sci.* 58, 669–673.
- Meltzoff, A. N., and Moore, M. K. (1997). Explaining facial imitation: a theoretical model. *Early Dev. Parent.* 6, 179–192. doi: 10.1002/(SICI)1099-0917(199709/12)6:3<179::AID-EDP157>3.0.CO;2-R
- Miller, N. E., and Dollard, J. (1941). *Social Learning and Imitation*. New Haven, CT: Yale University Press.

- Miura, K., Yoshikawa, Y., and Asada, M. (2012). Vowel acquisition based on an auto-mirroring bias with a less imitative caregiver. *Adv. Robot.* 26, 23–44. doi: 10.1163/016918611X607347
- Nagai, Y., Kawai, Y., and Asada, M. (2011). “Emergence of mirror neuron system: immature vision leads to self-other correspondence,” in *Development and Learning (ICDL), 2011 IEEE International Conference on*, (Frankfurt), 1–6.
- Perez-Orive, J., Mazor, O., Turner, G. C., Cassenaer, S., Wilson, R. I., and Laurent, G. (2002). Oscillations and sparsening of odor representations in the mushroom body. *Science* 297, 359–365. doi: 10.1126/science.1070502
- Prather, J. F., Peters, S., Nowicki, S., and R., M. (2008). Precise auditory-vocal mirroring in neurons for learned vocal communication. *Nature* 451, 305–310. doi: 10.1038/nature06492
- Reber, R., Winkielman, P., and Schwarz, N. (1998). Effects of perceptual fluency on affective judgments. *Psychol. Sci.* 9, 45–48. doi: 10.1111/1467-9280.00008
- Redgrave, P., and Gurney, K. (2006). The short-latency dopamine signal: a role in discovering novel actions? *Nat. Rev. Neurosci.* 7, 967–975. doi: 10.1038/nrn2022
- Rizzolatti, G., and Arbib, M. A. (1998). Language within our grasp. *Trends Neurosci.* 21, 188–194. doi: 10.1016/S0166-2236(98)01260-0
- Rizzolatti, G., and Craighero, L. (2004). The mirror-neuron system. *Annu. Rev. Neurosci.* 27, 169–192. doi: 10.1146/annurev.neuro.27.070203.144230
- Schmidhuber, J. (2009). “Driven by compression progress: a simple principle explains essential aspects of subjective beauty, novelty, surprise, interestingness, attention, curiosity, creativity, art, science, music, jokes,” in *Anticipatory Behavior in Adaptive Learning Systems*, (Springer), 48–76.
- Shepherd, S. V., Klein, J. T., Deaner, R. O., and Platt, M. L. (2009). Mirroring of attention by neurons in macaque parietal cortex. *Proc. Natl. Acad. Sci. U.S.A.* 106, 9489–9494. doi: 10.1073/pnas.0900419106
- Simoncelli, E. P., and Olshausen, B. A. (2001). Natural image statistics and neural representation. *Annu. Rev. Neurosci.* 24, 1193–1216. doi: 10.1146/annurev.neuro.24.1.1193
- Smith, E. C., and Lewicki, M. S. (2006). Efficient auditory coding. *Nature* 439, 978–982. doi: 10.1038/nature04485
- Sutton, R. S., and Barto, A. G. (1998). *Reinforcement Learning: An Introduction*. Cambridge, MA: Cambridge University Press.
- Thorpe, W. (1963). *Learning and Instinct in Animals*. 2nd Edn. Cambridge, MA: Harvard University Press.
- Triesch, J., Jasso, H., and Deák, G. O. (2007). Emergence of mirror neurons in a model of gaze following. *Adapt. Behav.* 14, 149–165. doi: 10.1177/1059712307078654
- Troyer, T. W., and Doupe, A. J. (2000). An associational model of birdsong sensorimotor learning I. Efference copy and the learning of song syllables. *J. Neurophysiol.* 84, 1204–1223.
- Zhao, Y., Rothkopf, C. A., Triesch, J., and Shi, B. E. (2012). “A unified model of the joint development of disparity selectivity and vergence control,” In *Development and Learning (ICDL), 2012 IEEE International Conference on*, (San Diego, CA), 1–6.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 11 July 2013; accepted: 10 October 2013; published online: 05 November 2013.

Citation: Triesch J (2013) Imitation learning based on an intrinsic motivation mechanism for efficient coding. *Front. Psychol.* 4:800. doi: 10.3389/fpsyg.2013.00800
This article was submitted to Cognitive Science, a section of the journal *Frontiers in Psychology*.

Copyright © 2013 Triesch. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Self-organization of early vocal development in infants and machines: the role of intrinsic motivation

Clément Moulin-Frier*, Sao M. Nguyen and Pierre-Yves Oudeyer

Flowers Team, Institut National de Recherche en Informatique et en Automatique / ENSTA-ParisTech, Bordeaux, France

Edited by:

Tom Stafford, University of Sheffield, UK

Reviewed by:

Minoru Asada, Osaka University, Japan

Ian Howard, University of Plymouth, UK

***Correspondence:**

Clément Moulin-Frier, Flowers Team, Institut national de recherche en informatique et en automatique / ENSTA-ParisTech, Bordeaux Sud-Ouest, 200 Avenue de la Vieille Tour, 33 405 Talence, France
e-mail: clement.moulinfrier@gmail.com

We bridge the gap between two issues in infant development: vocal development and intrinsic motivation. We propose and experimentally test the hypothesis that general mechanisms of intrinsically motivated spontaneous exploration, also called curiosity-driven learning, can self-organize developmental stages during early vocal learning. We introduce a computational model of intrinsically motivated vocal exploration, which allows the learner to autonomously structure its own vocal experiments, and thus its own learning schedule, through a drive to maximize competence progress. This model relies on a physical model of the vocal tract, the auditory system and the agent's motor control as well as vocalizations of social peers. We present computational experiments that show how such a mechanism can explain the adaptive transition from vocal self-exploration with little influence from the speech environment, to a later stage where vocal exploration becomes influenced by vocalizations of peers. Within the initial self-exploration phase, we show that a sequence of vocal production stages self-organizes, and shares properties with data from infant developmental psychology: the vocal learner first discovers how to control phonation, then focuses on vocal variations of unarticulated sounds, and finally automatically discovers and focuses on babbling with articulated proto-syllables. As the vocal learner becomes more proficient at producing complex sounds, imitating vocalizations of peers starts to provide high learning progress explaining an automatic shift from self-exploration to vocal imitation.

Keywords: vocal development, intrinsic motivation, curiosity-driven learning, imitation, self-organization, interactive learning, goal babbling

1. INTRODUCTION

1.1. VOCAL DEVELOPMENT AND INTRINSIC MOTIVATION

Early on, babies seem to explore vocalizations as if it was a game in itself, as reported by Oller (2000) who cites two studies from the nineteenth century:

"[At] 3 months were heard, for the first time, the loud and high crowing sounds, uttered by the child spontaneously, [...] the child seemed to take pleasure in making sounds." (Sigismund, 1971)

"[He] first made the sound *mm* spontaneously by blowing noisily with closed lips. This amused [him] and was a discovery for [him]."¹ (Taine, 1971)

Such play with his vocal tract, where the baby discovers the sounds he can make, echoes other forms of body play, such as exploration of arm movements or how he can touch, grasp, mouth or throw objects. The concept of *intrinsic motivation* has been proposed in psychology to account for such spontaneous exploration

(Berlyne, 1954; Deci and Ryan, 1985; Csikszentmihalyi, 1997; Ryan and Deci, 2000; Gottlieb et al., 2013):

"Intrinsic motivation is defined as the doing of an activity for its inherent satisfaction rather than for some separable consequence. When intrinsically motivated, a person is moved to act for the fun or challenge entailed rather than because of external products, pressures or reward." (Ryan and Deci, 2000)

Intrinsic motivation refers to a mechanism pushing individuals to select and engage in activities for their own sake because they are inherently interesting (in opposition to *extrinsic motivation*, which refers to doing something because it leads to a separable outcome). A key idea of recent approaches to intrinsic motivation is that *learning progress* in sensorimotor activities can generate intrinsic rewards in and for itself, and drive such spontaneous exploration (Gottlieb et al., 2013). Learning progress refers to the infant's improvement of his predictions or control over activity they practice, which can also be described as reduction of uncertainty (Friston et al., 2012).

Although spontaneous vocal exploration is an identified phenomenon, occurring in the early stages of infant development, the specific mechanisms of such exploration and the role of intrinsic motivation for the *structuration* of early vocal development has not received much attention so far to our knowledge. We propose

¹We have changed the gender of the subject to a male in this quotation, in order to follow the convention of the present article. Throughout this paper, we will use "he" for an infant, "she" for a caregiver (e.g., the mother) and "it" for a learning agent (the model).

that mechanisms of intrinsically motivated spontaneous exploration, which we also refer to as curiosity-driven learning, play an important role in speech acquisition, by driving the infant to follow a self-organized developmental sequence which will allow him to progressively learn to control his vocal tract. This is to our knowledge a largely unexplored hypothesis. The goal of this article is to formalize in detail this hypothesis and study general properties of such mechanisms in computer experiments.

Several computational models of speech development, where speech acquisition is organized along a developmental pathway, have been elaborated so far. They have shown how such stage-like organization can ease the acquisition of complex realistic speech skills.

The DIVA model (Guenther et al., 1998; Guenther, 2006), as well as Kröger's model (Kröger et al., 2009), propose architectures partly inspired by neurolinguistics. They involve two learning phases. The first one is analogous to infant babbling and corresponds to semi-random articulator movements producing auditory and somatosensory feedbacks. This is used to tune the correspondences between representation maps within a neural network. In the second phase, the vocal learner is presented with external speech sounds analogous to an ambient language and learns how to produce them adequately. The Elija model (Howard and Messum, 2011) also distinguishes several learning phases. In the first phase of exploration, the agent is driven by a reward function, including intrinsic rewards such as sound salience and diversity, as well as articulatory effort. Various parameterizations of this reward function allows the model to produce vocalizations in line with Oller's vocal developmental stages of infants. In a subsequent phase, the sounds produced by the model attract the attention of a caregiver, providing an external reinforcement signal. Other models also use a reinforcement signal, either from human listeners [social reinforcement (Warlaumont, 2012, 2013b)] or based on sound saliency [intrinsic reinforcement (Warlaumont, 2013a)], and show how this can influence a spiking neural network to produce canonical syllables. Such computational models of speech acquisition pre-determine the global ordering and timing of learning experiences, which amounts to preprogramming the developmental sequence. Understanding how a vocal developmental sequence can be formed is still a major mystery to solve, and this article attempts a first step in this direction.

We build on recent models of skill learning in other modalities (e.g., locomotion or object manipulation), where it was shown that mechanisms of intrinsically motivated learning can self-organize developmental pathways, adaptively guiding exploration and learning in high-dimensional sensorimotor spaces, involving highly redundant and non-linear mappings (Oudeyer et al., 2007; Baranes and Oudeyer, 2013; Gottlieb et al., 2013; Oudeyer et al., 2013). Such models concretely formalize concepts of intrinsic motivation described in the psychology literature into algorithmic architectures that can be experimented in computers and robots (Schmidhuber, 1991; Barto et al., 2004; Oudeyer and Kaplan, 2007; Baldassarre, 2011). Detailed discussions of the engineering aspects of such intrinsic motivation mechanisms, casted in the statistical framework of active learning, have been recently published and showed their algorithmic

efficiency to learn sensorimotor coordination skills in redundant non-linear high-dimensional mappings (Baldassarre and Mirolli, 2013; Baranes and Oudeyer, 2013; Srivastava et al., 2013).

Indeed, transposed in curiosity-driven learning machines (Schmidhuber, 1991; Barto et al., 2004; Schembri et al., 2007; Hart, 2009; Merrick and Maher, 2009; Schmidhuber, 2010; Stout and Barto, 2010) and robots (Oudeyer et al., 2007; Baranes and Oudeyer, 2013), these developmental mechanisms have been shown to yield highly efficient learning of inverse models in high-dimensional redundant sensorimotor spaces (Baranes and Oudeyer, 2010, 2013). These spaces share many mathematical properties with vocal spaces. Efficient versions of such mechanisms are based on the active choice of learning experiments that maximize learning progress, e.g., improvement of predictions or of competences to reach goals (Schmidhuber, 1991; Oudeyer and Kaplan, 2007; Oudeyer et al., 2007; Baranes and Oudeyer, 2013; Srivastava et al., 2013). Such learning experiments are called "progress niches" (Oudeyer et al., 2007).

Yet, beyond pure considerations of learning efficiency, exploration driven by intrinsic rewards measuring learning progress was also shown to self-organize structured developmental pathways, both behaviorally and cognitively. Indeed, such mechanisms automatically drive the system to explore and learn first easy skills, and then progressively explore skills of increasing complexity (Oudeyer et al., 2007). They have been shown to generate automatically behavioral and cognitive developmental structures and have been analyzed in relation to their similarities with infant development (Oudeyer and Kaplan, 2006; Kaplan and Oudeyer, 2007a; Oudeyer et al., 2007; Moulin-Frier and Oudeyer, 2012). For example, in the Playground Experiment, a curiosity-driven learning robot was shown to self-organize its own learning experiences into a sequence of behavioral and cognitive stages where it spontaneously acquired various affordances and skills of increasing complexity (Oudeyer et al., 2007). It was also shown how it could discover and focus on elementary vocal interaction with a peer as a spontaneous consequence of its general drive to explore situations where it can improve its predictions (Oudeyer and Kaplan, 2006). Focusing on vocal interactions was thus explained as a special case of focusing on an activity that provides learning progress (i.e., a particular progress niche). This therefore allowed to generate some novel hypotheses to explain infant development, from the behavioral (Oudeyer and Kaplan, 2006), cognitive (Kaplan and Oudeyer, 2007a), or brain circuitry (Kaplan and Oudeyer, 2007b) perspectives [see Gottlieb et al. (2013) for a review on these novel perspectives]. Intrinsically motivated spontaneous learning has also been combined with mechanisms of imitation learning within the SGIM-ACTS architecture, as detailed in Nguyen and Oudeyer (2012). In this model, formulated within the framework of strategic learning (Lopes and Oudeyer, 2012), a hierarchical active learning architecture allows an interactive learning agent to choose by itself when to explore autonomously, and when, what and who to imitate, based on measures of competence progress.

Although intrinsic motivation and socially guided learning have already been considered in computational models specifically studying speech acquisition, to our knowledge, they have so far been considered as two distinct learning phases with a

hard-coded switch between them (e.g., Guenther et al., 1998; Guenther, 2006; Kröger et al., 2009; Howard and Messum, 2011). In other words, the existence of distinct developmental stages was presupposed in these models. In contrast, these distinct learning phases emerge from the Playground Experiment, even though only a simplistic vocal system was considered (only pitch and duration were controlled, and no physical model of the vocal tract was used; modeling of speech acquisition *per se* was not the focus of this study).

Our main contribution in this paper is to show how mechanisms of intrinsically motivated exploration applied on a realistic articulatory-auditory system self-organizes autonomously into coherent *vocal* developmental sequences. This follows the approach of our previous works (Moulin-Frier and Oudeyer, 2012, 2013a,b), which were preliminary studies limited to vowel production and focusing only on autonomous learning, i.e., without considering a surrounding ambient language.

In such a conceptual framework, developmental structures are neither learnt from “tabula rasa” nor a pre-determined result of an innate “program”: they self-organize out of the dynamic interaction between constrained cognitive mechanisms (including curiosity, learning, and abstraction), the morphological properties of the body, and the physical and social environment which itself is constrained and ordered by the developmental level of the organism (Thelen and Smith, 1996; Oudeyer et al., 2007). Thus, the approach we take can be viewed as an instantiation of the concept of epigenesis, in the sense proposed by (Gottlieb, 1991).

The study of such a dynamical systems approach, where curiosity-driven learning is an important force, can take ample advantage of computer modeling as a research tool. Here in particular, it can help to understand better the dynamics underlying early vocal development, and in particular understand what are the mechanisms which generate the developmental sequence(s) in vocal productions and capabilities observed in infants. In particular, it can help to understand what is the precise role of intrinsic motivation.

In the next sections of this introduction, we summarize properties of vocal development during the first year and describe the general principles of the computational model we study in this article.

1.2. DEVELOPMENT OF VOCALIZATIONS

Despite inter-individual variations in infant vocal development (e.g., Vihman et al., 1986), strong regularities in the global structuration of vocal development are identified (Oller, 2000; Kuhl, 2004). In this article, we adopt the view from Oller (2000) as well as Kuhl (2004). **Figure 1** schematizes this vocal development during the first year of infant. It can be summarized as follows. First, until the age of approximately 3 months, an infant produces non-speech sounds like squeals, growls and yeals. During this period, he seems to learn to control infrastructural speech properties, e.g., phonation and primitive articulation (Oller, 2000). Then, from 3 to 7 months, he begins to produce vowel-like sounds (or quasi-vowels) while he probably learns to control his vocal tract resonances. At 7 months, canonical babbling emerges where well-timed sequences of proto-syllables are mastered. But it is

only around the age of 10 months that infant vocal productions become more influenced by the ambient language, leading to first word productions around 1 year of age.

Two features of this developmental sketch are particularly salient.

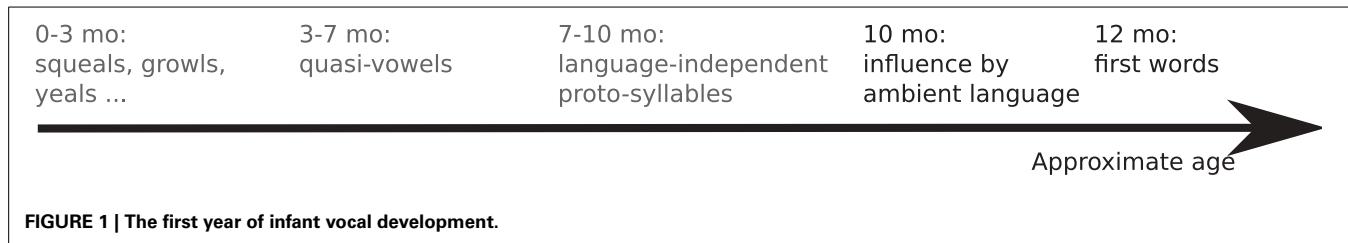
- Infants seem to first play with their vocal tracts in a relatively language-independent way, and then are progressively influenced by the ambient speech sounds.
- In the initial phase, when sounds produced by their peers influence little their vocalizations, infants seem to learn skills of increasing complexity: normal phonation, then quasi-vowels and finally proto-syllables. According to Oller (2000), such a sequence displays a so-called natural, or logical hierarchy. For example, it is impossible to master quasi-vowel production without previously mastering normal phonation.

1.3. A COMPUTATIONAL MODEL OF VOCAL DEVELOPMENT

To articulate hypotheses about the possible roles of intrinsic motivation in the first year of vocal development, we build here a computational model of an intrinsically motivated vocalizing agent, in contact with vocalizations of peers. In the model, an individual speech learner has the following characteristics, described in detail in next sections:

- It embeds a realistic model of a human vocal tract: the articulatory synthesizer used in the DIVA model (Guenther et al., 2006). This model provides the way to produce sequences of vocal commands and to compute corresponding sequences of acoustic features, both in multi-dimensional continuous domains.
- It embeds a dynamical model for producing motions of the vocal tract, based on an over-damped spring-mass model. This model describes dynamical aspects such as co-articulation in sequences of vocal targets.
- It is able to iteratively learn a probabilistic sensorimotor model of the articulatory-auditory relationships according to its own experience with the vocal tract model. Because the sensorimotor learning is iterative during the life time of the agent, it will first be inefficient at using this model for control, and then progresses by learning from its own experience.
- It is equipped with an intrinsically motivated exploration mechanism, which allows it to generate and select its own auditory goal sequences. Such mechanism includes a capability to empirically measure its own competence progress to reach sequences of goals. Then, an action selection system stochastically self-selects target goals that maximize competence progress.
- It is able to hear sounds of a simulated ambient language, and its intrinsic motivation system is also used to decide whether to self-explore self-generated auditory goals, or to try to emulate adult sounds. This choice is also based on a measure of competence progress for each strategy.

Then, we present experiments allowing us to study how the developmental structuration of early vocal exploration could be self-organized in an intrinsically motivated speech learner, under



the influence of sounds in the environment and constrained by the physical properties of the sensorimotor system.

In a first series of experiments, we consider a speech learner who is not exposed to external speech sounds. This allows the study of the role of intrinsic motivation independently of any social influence. We show how a cognitive architecture for intrinsically motivated autonomous exploration (SAGG-RIAC; Baranes and Oudeyer, 2013; Moulin-Frier and Oudeyer, 2013a), applied to learning to control an articulatory synthesizer (i.e., a vocal tract model able to produce speech sounds from articulatory configurations), can self-organize coherent vocal developmental sequences. This work extends preliminary studies (Moulin-Frier and Oudeyer, 2012, 2013a,b) through the use of a different vocal tract model and a more complex model of motion control dynamics with an overdamped spring-mass dynamical system, providing the agent with a more realistic and powerful mechanism to produce (un)articulated sounds.

In a second series of experiments, the speech learner is exposed to speech sounds from its environment. The cognitive architecture is extended to strategic interactive intrinsically motivated learning (SGIM-ACTS; Nguyen and Oudeyer, 2012), where intrinsic motivation is also used by the learner to decide when to self-explore and when to try to imitate sounds in the environment. In the present study, we suppose that the sounds of the adult are directly imitable (we do not account for the pitch and formant differences between infants and adults for instance). We show that the system first focuses on self-exploration of vocalization. It later on shifts to vocal imitation, which then influences its vocal learning in ways that are specific to the speech environment. Yet, in this paper, we do not study the social interaction aspect of the teacher and, in particular, we do not model the behavior of the adult in response to the learner behavior.

Our aim is to study how important aspects of infant vocal development in the first year of life, described in the previous section, could be explained by the interaction between these building blocks: an intrinsic motivation system, a dynamic motor system associated to morphological and physiological constraints, an imitation system and a system for learning a sensorimotor model out of physical experiments. We will show that competence progress based autonomous exploration is able to provide a unified explanation for both the tendency to produce vocalizations of increasing complexity and the progressive influence of the ambient adult sounds. Imitating adult sounds becomes interesting for the speech learner only when basic speech production principles have been previously mastered. Contrarily to existing models of speech acquisition we described so far, our aim is not to reproduce infant vocalizations in a phonetically detailed manner, but rather to suggest an hypothesis about how a succession

of distinct developmental stages can self-organize autonomously. Howard and Messum's model (Howard and Messum, 2011) for example, shows how distinct parameterizations of an intrinsic reward function can enable a vocal agent to discover several type of sounds coherent with observed developmental stages in infants. These parameterizations however, are hard-coded. In contrast, our model is not designed to reproduce precisely infant vocalizations within distinct vocalization stages, but rather to understand how the *transition* from one stage to another can be explained by a drive to maximize the competence progress to reach self-generated or ambient auditory goals. In consequence, the switch from self-generated auditory goals to the imitation of adult sounds is not hard-coded in our model, but emerges as a by-product of the drive to focus on progress niches.

2. MODEL

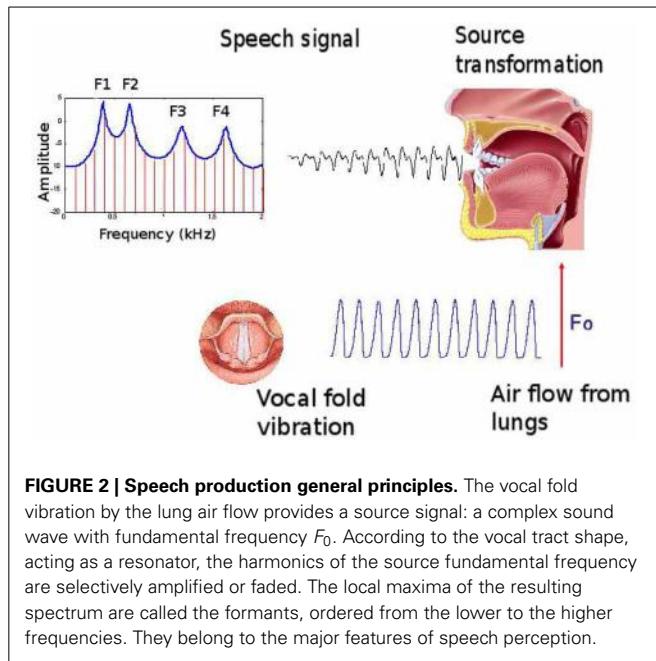
In this section, we describe the models that we use for the vocal tract and auditory signals. We describe the learning of the internal model of the sensorimotor mapping, and the intrinsic motivation mechanism which allows the learner to decide adaptively which vocalization to experiment at given moments during its development, and whether to do so through self-exploration or through imitation of external sounds.

2.1. SENSORIMOTOR SYSTEM

2.1.1. Vocal tract and auditory system

Our computational model involves the articulatory synthesizer of the DIVA model described in Guenther et al. (2006)² based on Maeda's model (Maeda, 1989). Without going into technical details, the model corresponds to a computational approximation of the general speech production principles illustrated in Figure 2. The model receives 13 articulatory parameters as input. The first 10 are from a principal component analysis (PCA) performed on sagittal contours of images of the vocal tract of a human speaker, allowing to reconstruct the sagittal contour of the vocal tract from a 10-dimensional vector. The effect of the 10 articulatory parameters from the PCA on the vocal tract shape is displayed Figure 3. In this study, we will only use the 7 first parameters (the effect of the others on the vocal tract shape is negligible), fixing the 3 last in the neutral position (value 0 in the software). Through an area function, associating sections of the vocal tract with their respective area, the model can compute the 3 first formants of the resulted signal if phonation occurs.

²Available online at <http://www.bu.edu/speechlab/software/diva-source-code>. DIVA is a complete neurocomputational model of speech acquisition, in which we only use the synthesizer computing the articulatory-to-auditory function.

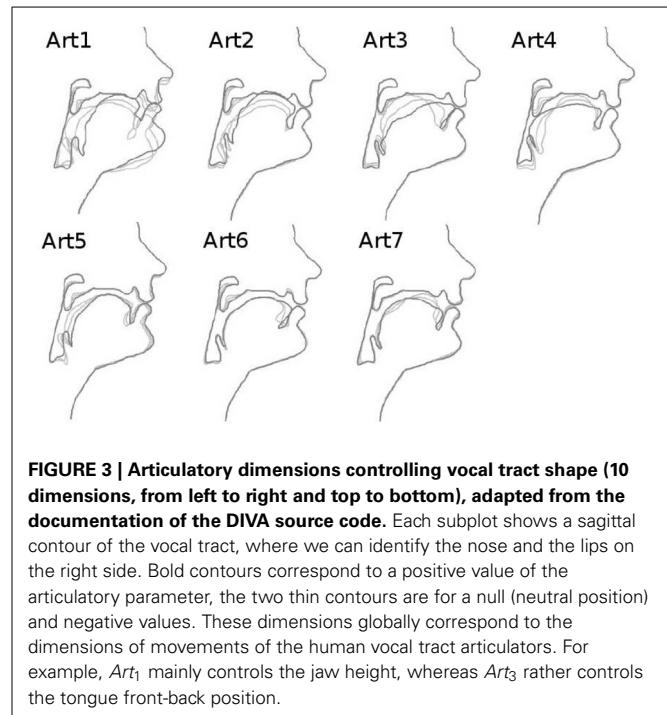


Phonation is controlled through the 3 last parameters: glottal pressure controlling the intensity of the signal (from quiet to loud), voicing controlling the voice (from voiceless to voiced) and pitch controlling the tone (from low-pitched to high-pitched). It is then able to compute the formants of the signal (among other auditory and somato-sensory features) through the area function. In this study, we only use the glottal pressure and voicing parameters. In addition to the 7 articulatory parameters from the PCA, a vocal command is therefore defined by a 9-dimensional vector. From the vocal command, the synthesizer computes the auditory and somatosensory consequences of the motor command, thus approximating the speech production principles of **Figure 2**.

On the perception side of our model, we use the first two formants of the signal, F_1 and F_2 , approximately scaled between -1 and 1 . We also define a third parameter I which measures the intensity (or phonation level) of the auditory outcome. I is supposed to be 0 when the agent perceives no sound, and 1 when it perceives a sound. Technically, $I = 1$ if and only if two conditions are checked: (1) both pressure and voicing parameters are above a fixed threshold (null value) and (2) the vocal tract is not closed (i.e., the area function is positive everywhere). In human speech indeed, the formants are not measurable when phonation is under a certain threshold. We model this by setting that when $I = 0$, the formants do not exist anymore and are set to 0 . This drastic simplification is yet arguable in term of realism, but what we want to model here is the fact that no control of the formant values can be learnt when no phonation occurs.

2.1.2. Dynamical properties

Speech production and perception are dynamical processes and the principles of **Figure 2** have to be extended with this respect. Humans control their vocal tract by variations in muscle activations during a vocalization, modulating the produced sound in a complex way. Closure or opening movements during a particular



vocalization, coupled with variations in phonation level, are able to generate a wide variety of modulated sounds. We thus define a vocalization as a trajectory of the 9 motor parameters over time, lasting 800 ms, from which the articulatory synthesizer is able to compute the corresponding trajectories in the auditory space (i.e., trajectories in the 3-dimensional space of F_1 , F_2 , and I). The agent is able to control this trajectory by setting 2 commands for each articulator: one from 0 to 250 ms, the other one from 250 to 800 ms. Then, the motor system is modeled as an over-damped spring-mass system driven by the following second-order dynamical equation:

$$\ddot{x} + 2\zeta\omega_0\dot{x} + \omega_0^2(x - m) = 0, \quad (1)$$

where x is a motor parameter, and m is the command for that motor parameter. ζ is set to 1.01 , ensuring that the system is over-damped (no oscillation), and ω_0 to $\frac{2\pi}{0.8}$ (0.8 being the duration of the vocalization in seconds). Thus, the agent's policy for a vocalization is defined by two vectors m_1 and m_2 (one for each command) of 9 real values each (one for each motor parameter). The policy space is 18-dimensional. The first command is applied for the beginning of the vocalization to 250 ms, the second one from 250 to 800 ms.

Figure 4A illustrates the process by showing a typical syllabic vocalization. In this illustrative example, the controlled articulators are the first and third articulators of **Figure 3** (roughly controlling the jaw height and the tongue front/back dimensions), as well as pressure and voicing. The two last ones are set to 0.5 and 0.7 , respectively, for both commands, to allow phonation to occur. The "jaw parameter" (art_1 on the figure) is set to 2.0 (jaw closed) for the first command and to -3.0 for the second one (jaw open). We observe that these commands, quite far from the

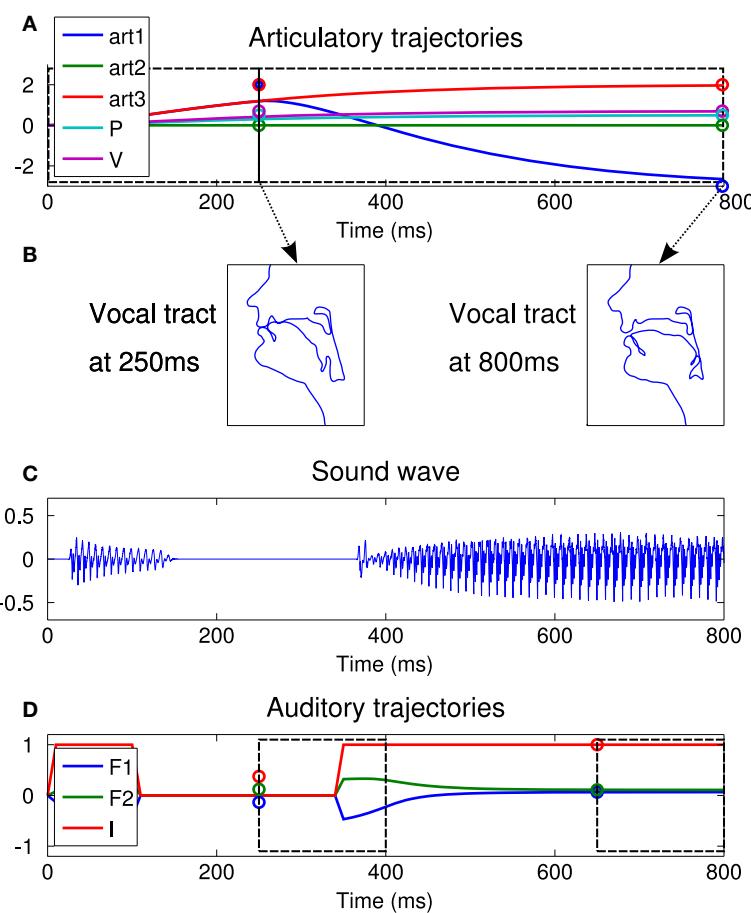


FIGURE 4 | An illustrative vocalization example. (A) Articulatory trajectories of 5 articulators during the 800 ms of the vocalization (4 articulators, from $art4$ to $art7$ are not plotted for the sake of readability but display the same trajectory as $art2$). Circles at 250 and 800 ms represent the values of the first and second commands, respectively, for each trajectory. The first commands are active from 0 to 250 ms and second ones from 250 to 800 ms, as represented by dotted black boxes. The trajectories are computed by the second order dynamical Equation (1), starting in a neutral

position (all articulators set to 0). (B) Resulting vocal tract shapes at the end of each command, i.e., at 250 and 800 ms. Each subplot displays a sagittal view with the nose and the lips on the left side. The tongue is therefore to the right of the lower lip. (C) Sound wave resulting from the vocalization. (D) Trajectories of the 3 auditory parameters, the intensity I and the two first formants $F1$ and $F2$. Dotted black boxes represent the two perception time windows. The agent perceives the mean value of the auditory parameters in each time window, represented by the circles at 250 and 650 ms.

neutral position, are not completely reached by the motor system. This is due to the particular dynamics of the system, defined with ζ and ω_0 in the dynamical system. For the third articulator ($art3$), the commands are both at 2.0. We observe that, whereas the value 2.0 cannot be achieved completely at 250 ms, it can however be reached before the end of the vocalization.

This motor system implies interaction between the two commands, i.e., a form of co-articulation. Indeed, a given motor configuration may sometimes be harder to reach if it is set as the first command, because time allocated to reach the first command is less than for the second command. Reversely, some movements may be harder to control in the second command because the final articulator positions will depend both on the first and the second commands (e.g., it is harder to reach the value -3.0 for the second command if the first command is set to 2.0, than if the first command is set to -3.0 , as seen in the example of Figure 4).

These characteristics are the results of modeling speech production as a damped spring-mass system (Equation 1), which is a common practice in the literature (Markey, 1994; Boersma, 1998; Howard and Messum, 2011).

Figure 4B shows the resulting vocal tract shape at the end of the 2 commands (i.e., at 250 ms and at 800 ms). We observe that the vocal tract is closed at the end of the first command, open at the end of the second one.

Figure 4C shows the resulting sound. We observe that there is no sound during vocal tract closure.

Figure 4D shows the resulting trajectories of auditory parameters. In our experiments, we model the auditory perception of the agent of its own vocalization as the mean value of each parameter I , $F1$, and $F2$ in two different time windows lasting 150 ms: the first one from 250 to 400 ms, the second one from 650 to 800 ms. The auditory representation of a vocalization is therefore a 6-dimensional vector $[I_{(1)}, I_{(2)}, F1_{(1)}, F1_{(2)}, F2_{(1)}, F2_{(2)}]$.

Perceived auditory values are represented by circles on **Figure 4D**. Note that the agent does not have any perception of what happens before 250 ms, and that $I_{(1)}$ and $I_{(2)}$ can take continuous values in [0, 1] due to the averaging in a given perception time window. We will refer to the perceived “phone” of a given command for the perception occurring around the end of that command, although such an association will not be assumed in the internal sensorimotor model of the agent. Indeed, this sensorimotor system has the interesting property that the perceptions in both time windows depend on both motor commands. In the example of **Figure 4**, the perception for the first command, i.e., the mean auditory values between 250 and 400 ms, would not be the same if the second motor command did not cause the vocal tract opening.

2.1.3. Vocalization classification

We define three types of phones, according to the value of I for a given command. In this description, we use common concepts like vowels or consonants to make an analogy with the human types of phones, although this analogy is limited.

- Those where $I > 0.9$, i.e., phonation occurs during almost all the 150 ms of perception around the end of the command. We call them *Vowels* (V).
- Those where $I < 0.1$, i.e., there is almost no phonation during the 150 ms of perception around the end of the command. We call them *None* (N).
- Those where $0.1 < I < 0.9$, i.e., phonation occurs partially during the 0.15 s of perception around the end of the command. This means that the phonation level I has switched during that period. This can be due either to a closure or opening of the vocal tract, or to variations in the pressure and voicing parameters. We call them *Consonants* (C), although they are sometimes more comparable to a sort of prosody (when due to a variation in the phonation level).

This classification will be used as a tool for the analysis of the results in section 3, but is never known by the agent (which only has access to the values of I , F_1 , and F_2).

Thus, each vocalization produced by the agent, belongs to the combination of 2 of these 3 types (because a vocalization corresponds to 2 commands), i.e., there are $3^2 = 9$ types of vocalizations: VV, VN, VC, NV, NN, NC, CV, CN, CC. An example of each type is given in the Appendix, section .

Then, we suggest to group these 9 types into 3 classes.

- The class *No Phonation* contains only NN: the agent has not produced an audible sound. This is due either to the fact the pressure and voicing motor variables have never been sufficiently high (not both positive, as explained in the description of the motor system) during the two 150 ms perception periods, or that the vocal tract was totally closed.
- The class *Unarticulated* contains VN, NV, CN, NC: the vocalization is not well-formed. Either the first or the second command produces a phone of type *None* ($I < 0.1$, see above).
- The class *Articulated* contains CV, VC, VV and CC: the vocalization is well-formed, in the sense that there is no *None* phone. Phonation is modulated in most cases (i.e., except in the rare

case where the two commands of a VV are very similar). Note that according to the definition of *consonants*, phonation necessarily occurs in both the perception time windows (see **Figure A1** in the Appendix).

It is important to note that the auditory values of these vocalization classes span subspaces of increasing complexity. Indeed, whereas various articulatory configurations belong to the *No Phonation* class, their associated auditory values are always null, inducing a 0-dimensional auditory subspace (i.e., a point). Regarding the *Unarticulated* class, the associated auditory values span a 3-dimensional subspace because at least one command produces a phone of type *None* (i.e., the corresponding auditory values are null). Finally, in the *Articulated* classes, the auditory values span the entire 6-dimensional auditory space. These properties will have important consequences for the learning of a sensorimotor model by the agent, as we will see.

2.2. INTERNAL SENSORIMOTOR MODEL

The sensorimotor internal model and the intrinsic motivation system which follow were firstly described in conference papers (Moulin-Frier and Oudeyer, 2013a,b) in a more general context where the goal was to compare various exploration strategies. In this paper, we use the active goal exploration strategy—analog to the SAGG-RIAC algorithm in Baranes and Oudeyer (2010, 2013).

During its life time, the agent iteratively updates an internal sensorimotor model by observing the auditory results of its vocal experiments. We denote motor commands M and sensory perceptions S . We call $f : M \rightarrow S$ the unknown function defining the physical properties of the environment (including the agent's body). When the agent produces a motor command $m \in M$, it then perceives $s = f(m) \in S$, modulo an environmental noise and sensorimotor constraints. In the sensorimotor system defined in the previous section, M is 18-dimensional and S is 6-dimensional. f corresponds to the transformation defined section 2.1 and illustrated **Figure 4**, and has a Gaussian noise with a standard deviation of 0.01. By collecting (m, s) pairs through vocal experiments, the agent learns the joint probability distribution defined over the entire sensorimotor space SM (therefore 24-dimensional). This distribution is encoded in a Gaussian Mixture Model (GMM) of 28 components, i.e., a weighted sum of 28 multivariate normal distributions³. Let us note GSM this GMM. It is learnt using an online version of the Expectation-Maximization (EM) algorithm (Dempster et al., 1977) proposed by Calinon (2009) where incoming data are considered incrementally. Each update is executed once each $sm_step (= 400)$ vocalizations are collected. GSM is thus refined incrementally during the agent life, updating each time a number sm_step of new (m, s) pairs are collected. Moreover, we adapted this online version of EM to introduce a *learning rate* parameter α which decreases logarithmically from 0.1 to 0.01 over time. α allows to set the relative weight of the new learning data with respect to the old ones.

This GMM internal model is used to solve the inverse problem of inferring motor commands $m \in M$ that allow the learner

³We empirically chose a number of components which is a suitable trade-off between learning capacity and computational complexity.

to reach a given auditory goal $s_g \in S$. From this sensorimotor model G_{SM} , the agent can compute the distribution of the motor variables knowing a given auditory goal to reach s_g , noted $G_{SM}(M | s_g)$. This is done by Bayesian inference on the joint distribution, and results in a new GMM over the motor variables M (see e.g., Calinon, 2009), from which the agent can sample configurations in M .

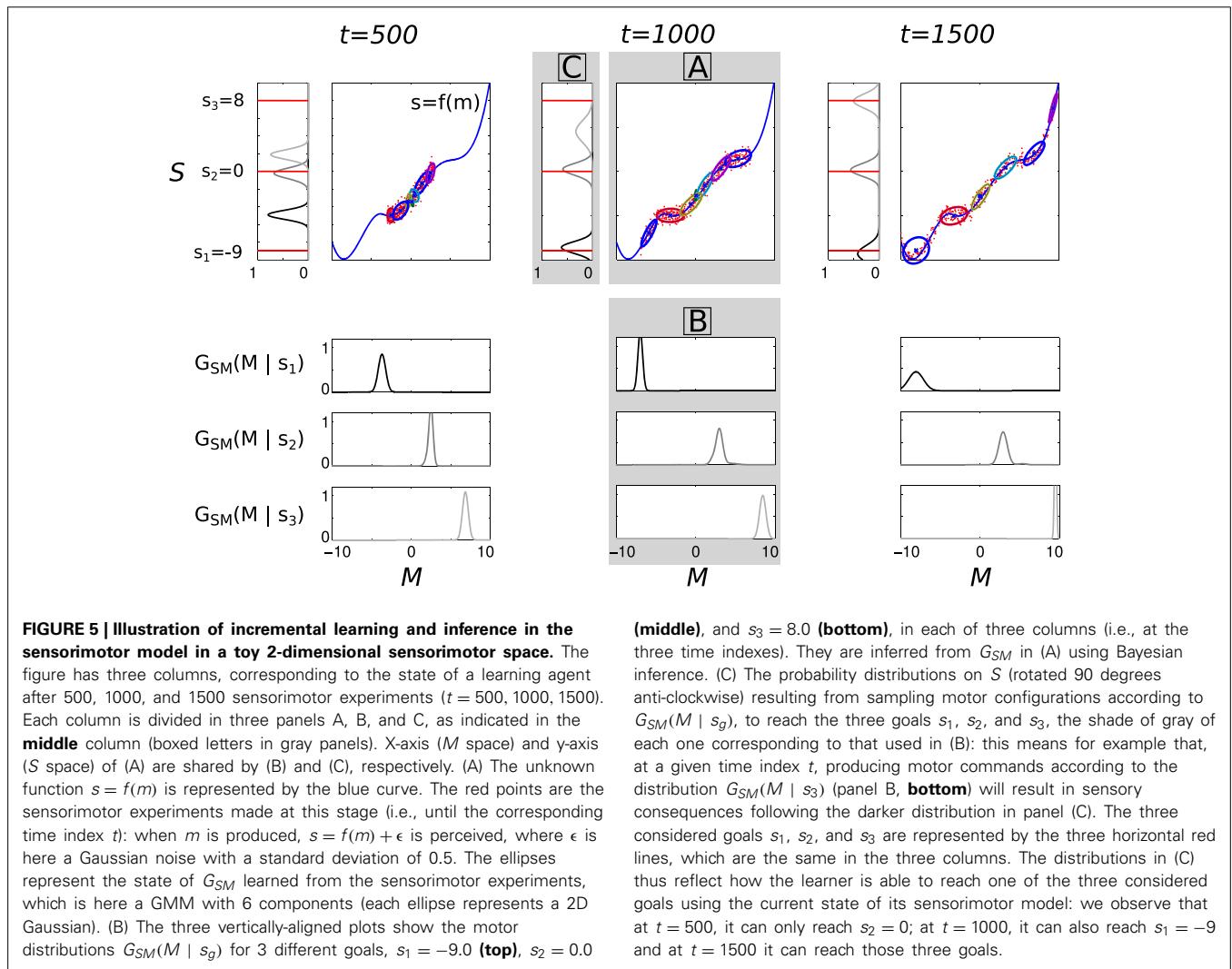
The whole process is illustrated **Figure 5**, on a toy example with mono-dimensional M and S . Given the current state of the sensorimotor model, the agent tries to achieve three goals, $s_1 = -9$, $s_2 = 0$, and $s_3 = 8$, i.e., three points in S (how the agent is going to self-generate such goals with intrinsic motivation will be explained below). At the beginning of the life time, the model is very poor at finding a good solution because the GMM is trained with only a few data, not necessarily concentrated in the regions useful to achieve the goals. For example, at $t = 500$, the agent is only able to correctly reach $s_2 = 0$ but is inefficient at reaching $s_1 = -9$ and $s_3 = 8$, as shown by the distributions over S in the top left corner (rotated 90 degrees anti-clockwise). Then it becomes better and better while the agent produces new vocalizations, covering a larger part of the sensorimotor

space: at $t = 1500$, the agent is able to reach the three goals.

The sensorimotor system we specified in the previous section, however, involves a 24-dimensional sensorimotor space (18 articular dimensions and 6 auditory ones). Moreover, as we have already noted, the three vocalization classes we defined (*No Phonation*, *Unarticulated*, and *Articulated*) span subspaces of the 6-dimensional auditory space with increasing dimensionality. Learning an inverse model using GMMs with a fixed number of Gaussians is harder, i.e., requires more sensorimotor experiments, as the spanned auditory subspace is of higher dimensionality. Although we do not provide mathematical arguments to this claim in this paper, it seems clear that learning an inverse model to produce *No Phonation* requires fewer learning data than learning an inverse model to produce various *Articulated* vocalizations, because the range of sensory effect is much larger in the second case.

2.3. INTRINSICALLY MOTIVATED ACTIVE EXPLORATION

In order to provide training data to the sensorimotor model we just described, the agent autonomously and adaptively decides



which vocal experiments to make. The key idea is to self-generate and choose goals for which the learner predicts that experiments to reach these goals will lead to maximal competence progress.

The specific model we use in the first series of experiments (section 3.1) is a probabilistic version of the SAGG-RIAC architecture (Baranes and Oudeyer, 2010, 2013). This architecture was itself derived as a functional model (Oudeyer and Kaplan, 2007; Gottlieb et al., 2013) of theories in psychology (Berlyne, 1954; Deci and Ryan, 1985; Csikszentmihalyi, 1997; Ryan and Deci, 2000) which describe spontaneous exploration and curiosity in humans. It combines two principles: (1) goal babbling, also called goal exploration; (2) active learning driven by the maximization of empirically measured learning progress [which corresponds to the active goal strategy in Moulin-Frier and Oudeyer (2013a,b)]. In practice, the learner self-generates its own auditory goals in the sensory space S . One goal is here a sequence of two auditory targets encoded in a 6-dimensional vector $s_g = [I_{(1)}, I_{(2)}, F1_{(1)}, F1_{(2)}, F2_{(1)}, F2_{(2)}]$ (see section 2.1). For each goal, it uses the current sensorimotor estimation to infer a motor program $m \in M$ in order to reach that goal. Through the sensorimotor system, this produces a vocalization and the agent perceives the auditory outcome $s \in S$, hence a new (m, s) training data. Goals are selected stochastically so as to maximize the expected competence progress (i.e., the learner is

interested in goals where it predicts it can improve maximally its competence to reach them at a particular moment of its development). This allows the learner to avoid spending too much time on unreachable or trivial goals, and progressively explore self-generated goals/tasks of increasing complexity. As a consequence, the learner self-explores and learns only sub-parts of the sensorimotor space that are sufficient for reachable goals: this allows to leverage the redundancy of these spaces by building dense tubes of learning data only where it is necessary for control.

We define the competence c associated to a particular experiment (m, s) to reach the goal s_g as $c = \text{comp}(s_g, s) = e^{-\|s_g - s\|}$. This measure is in $[0, 1]$ and exponentially increases toward 1 when the Euclidean distance between the goal and the actual realization $s = f(m) + \epsilon$ tends to 0.

The measure of competence progress uses another GMM, G_{IM} , learnt using the classical version of EM on the recent goals and their associated competences. This GMM provides an interest distribution $G_{IM}(S)$ used to sample goals in the auditory space S maximizing the competence progress in the recent sensorimotor experiments of the agent. This was firstly formalized in Moulin-Frier and Oudeyer (2013a,b). In this paper, we provide a graphical explanation of the process in **Figure 6**.

Following all the previous definitions, we now consider that the agent possesses the following abilities:

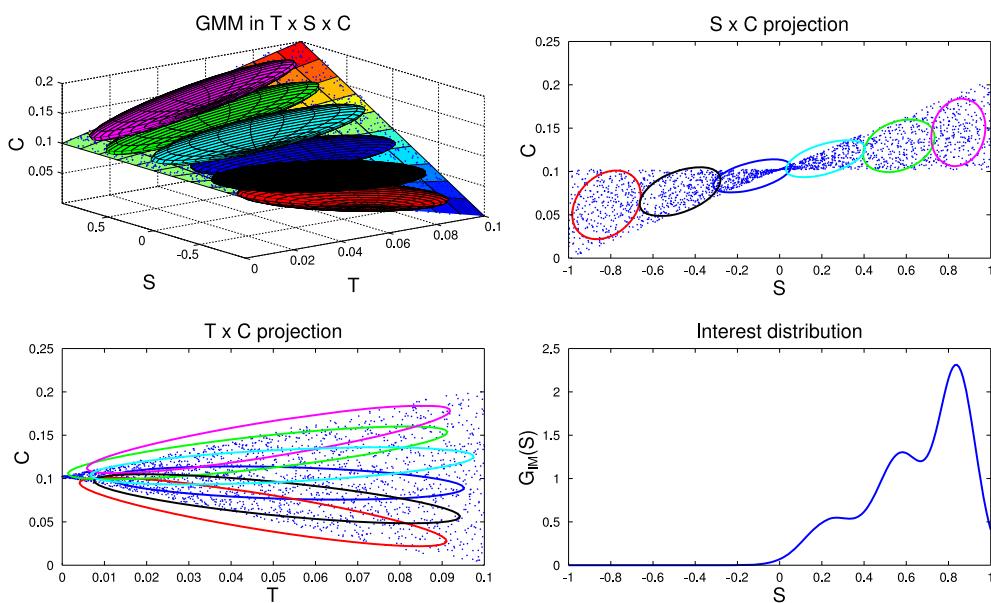


FIGURE 6 | Illustration of interest distribution computation. **Top-Left:** the recent history of competences of the agent, corresponding to blue points in the space $T \times S \times C$, where T is the space of recent time indexes (in \mathbb{R}^+), S the space of recently chosen goals s_g (mono-dimensional in this toy example) and C the space of recent competences of reaching those goals (in \mathbb{R}^+). For the sake of the illustration, the competence variations over time are here hand-defined (surf surface) and proportional to the values in S (increases for positive values, decreases for negative values). We train a GMM of 6 components, G_{IM} , to learn the joint distribution over $T \times S \times C$, represented by the six 3D ellipses. Projections of these ellipses are shown in 2D spaces $S \times C$ and $T \times C$ in the **Top-Right** and

Bottom-Left plots. To reflect the competence progress in this dataset, we then bias the weight of each Gaussian to favor those which display a higher competence progress, that we measure as the covariance between time and competence for each Gaussian (in the example the magenta ellipse shows the higher covariance in the **Bottom-Left** plot, then the green one, the sky blue one etc). We weight the Gaussians with a negative covariance between time T and competence C (blue, black, and red ellipses) with a negligible factor, such that they do not contribute to the mixture. Using Bayesian inference in this biased GMM, we finally compute the distribution over the goal space S , $G_{IM}(S)$, thus favoring regions of S displaying the highest competence progress (**Bottom-Right**).

- Producing a complex vocalization, sequencing two motor commands interpolated in a dynamical system. It is encoded by a 18-dimensional motor configuration $m \in M$.
- Perceiving the 6-dimensional auditory consequence $s = f(m) + \epsilon \in S$, computed by an articulatory synthesizer. f is unknown to the agent.
- Iteratively learning a sensorimotor model from lots of (m, s) pairs it collects by vocalizing through time. It is encoded in a GMM G_{SM} over the 24-dimensional sensorimotor space $M \times S$.
- Controlling its vocal tract to achieve a particular goal s_g . This is done by computing $G_{SM}(M | s_g)$, the distribution over the motor space M knowing a goal to achieve s_g .
- Actively choosing goals to reach in the sensory space S by learning an interest model G_{IM} in the recent history of experiences. By sampling in the interest distribution $G_{IM}(S)$, the agent favors goals in regions of S which maximizes the competence progress.

This agent is thus able to act at two different levels. At a high level, it chooses auditory goals to reach according to its interest model G_{IM} maximizing the competence progress. At a lower level, it attempts to reach those goals using Bayesian inference over its sensorimotor model G_{SM} , and incrementally refines this latter with its new experiences. The combination of both levels results in a self-exploration algorithm (**Algorithm 1**).

The agent starts in line 1 with no experience in vocalizing. Both GMMs have to be initialized in order to be used. To do this, the agent acquires a first set of (m, s) pairs, by sampling in M around the neutral values of the articulators (see **Figure 3**). Regarding the pressure and voicing motor parameters, we consider that the neutral value is at -0.25 , which leads to *no phonation* (recall that both these parameters have to be positive for phonation to occur, section 2.1). This models the fact that the agent does not phonate in its neutral configuration, and has at least to raise the pressure and voicing parameters to be able to do it. The agent then executes this first set of motor configurations (mostly not phonatory), observes the sensory consequences, and initializes G_{SM} with the corresponding (m, s) pairs using incremental EM. G_{IM} is initialized by setting the interest distribution $G_{IM}(S)$ to the distributions of the sounds it just produced with this first set of experiences. Thus, at the first iteration of the algorithm, the agent tries to achieve auditory goals corresponding to the

sounds it produced during the initialization phase. Then, in the subsequent iterations, the interest distribution $G_{IM}(S)$ reflects the competence progress measure, and is computed as explained above.

Line 3, the agent thus selects stochastically $s_g \in S$ with high interest values. Then it uses $G_{SM}(M | s_g)$ to sample a vocalization $m \in M$ to reach s_g (line 4). The execution of m will actually produce an auditory outcome s (line 5), and a competence measure to reach the goal, $c = comp(s_g, s)$, is computed (line 6). This allows it to update the sensorimotor model G_{SM} with the new (m, s) pairs (line 7). Finally, it updates the interest model G_{IM} (line 8) with the competence c to reach s_g

Algorithm 1 will be run and the results analyzed in section 3.1.

2.4. SOCIAL (OR IMITATION) SYSTEM

In language acquisition and vocalization, the social environment plays naturally an important role. Thus we consider an active speech learner that not only can self-explore its sensorimotor space, but can also learn by imitation. In a second series of experiments (section 3.2), we extend the previous model by integrating the previous learning algorithm in the SGIM-ACTS architecture, which has been proposed in Nguyen and Oudeyer (2012).

We consider here that the learning agent can use one of two learning strategies, which it chooses adaptively:

- explore autonomously with intrinsically motivated goal babbling, as described previously,
- or explore with imitation learning. We distinguish mimicry, in which the learner copies the policies of others without an appreciation of their purpose, from emulation, where the observer witnesses someone producing an outcome, but then employs its own policy repertoire to reproduce the outcome, as formalized in Whiten (2000); Call and Carpenter (2002); Nehaniv and Dautenhahn (2007); Lopes et al. (2010). As the learner a priori can not observe the vocal tract of the demonstrator, it can only emulate the demonstrator by trying to reproduce the auditory outcome observed, by using its own means, finding its own policy to reproduce the outcome. We consider that the demonstrator (the social peer) has a finite set of auditory outcomes, and every time the learner chooses to learn by social guidance, it chooses at random an auditory outcome among the set to emulate.

The learner can monitor the competence progress resulting from using each of the strategies. This measure is used to decide which strategy is the best progress niche at a given moment: a strategy is chosen with a probability directly depending on its associated expected competence progress. Thus, competence progress is used at two hierarchical levels of active learning, forming what is called strategic learning (Lopes and Oudeyer, 2012): at the higher-level, it is used to decide when to explore autonomously, and when to imitate; at the lower-level, if self-exploration is selected, it is used to decide which goal to self-explore (as in the previous model). Since competence progress is a non-stationary measure and is continuously re-evaluated, the individual *learns* to choose both the strategy

Algorithm 1 | Self-exploration with active goal babbling (stochastic SAGG-RIAC architecture).

```

1: initialise  $G_{SM}$  and  $G_{IM}$ 
2: while true do
3:    $s_g \sim G_{IM}(S)$ 
4:    $m \sim G_{SM}(M | s_g)$ 
5:    $s = f(m) + \epsilon$ 
6:    $c = comp(s_g, s)$ 
7:   update( $G_{SM}, (m, s)$ )
8:   update( $G_{IM}, (s_g, c)$ )
9: end while

```

$str \in \{autonomous_exploration, social_guidance\}$ and the auditory goals $s_g \in S$ to target, by choosing which combination enables highest competence progress.

For the particular implementation of SGIM-ACTS of this paper, we use the same formalism and implementation as in **Algorithm 1** and consider that the strategy is another choice made by the agent. This leads to **Algorithm 2**, where the interest model G_{IM} now learns an interest distribution as in section 2.3. The difference is that the space of interest is now the union of the strategy space $\{autonomous_exploration, social_guidance\}$ and the auditory space S . We call $StrS$ this new space $StrS = \{autonomous_exploration, social_guidance\} \times S$. Hence G_{IM} is a distribution over $StrS$ (**Algorithm 2**, line 3). If the self-exploration strategy is chosen ($str = autonomous_exploration$), the agent acts as in **Algorithm 2**. If the social guidance strategy is chosen ($str = social_guidance$, line 4), the learner then emulates an auditory demonstration $s_g \in S$ chosen randomly among the demonstration set of adult sounds (line 5), overwriting s_g of line 3. It then uses its sensorimotor model G_{SM} to choose a vocalization $m \in M$ to reach s_g , by drawing according to the distribution $G_{SM}(M | s_g)$ (line 7), as in the self-exploration strategy. The execution of m will produce an auditory outcome s (line 8), from which it updates its models G_{IM} and G_{SM} (lines 10 and 11).

Thus, this new exploration algorithm is augmented with yet another level of learning, allowing to choose between different exploration strategies. This strategy choice moreover uses the same mechanism as the choice of auditory goals, by means of the interest model G_{IM} .

Algorithm 2 will be run and the results analyzed in section 3.2.

3. RESULTS

The results of our experiments are presented in this section. We first run experiments where our agent learns in a pure self-exploration mode (**Algorithm 1**), without any social environment or sounds to imitate. In a second time, we introduce an auditory environment to study the influence of ambient language (**Algorithm 2**).

Algorithm 2 | Strategic active exploration (active goal babbling and imitation with stochastic SGIM-ACTS architecture).

```

1: Initialize  $G_{SM}$  and  $G_{IM}$ 
2: while true do
3:    $(str, s_g) \sim G_{IM}(StrS)$ 
4:   if ( $str = social\_guidance$ ) then
5:      $s_g \leftarrow$  random auditory demonstration from the ambient language
6:   end if
7:    $m \sim G_{SM}(M | s_g)$ 
8:    $s = f(m) + \epsilon$ 
9:    $c = comp(s_g, s)$ 
10:   $update(G_{SM}, (m, s))$ 
11:   $update(G_{IM}, (str, s_g, c))$ 
12: end while
```

3.1. EMERGENCE OF DEVELOPMENTAL SEQUENCES IN AUTONOMOUS VOCAL EXPLORATION

We ran 9 independent simulations of **Algorithm 1** with the same parameters but different random seeds, of 240,000 vocalizations each⁴. Most of these 9 simulations display the formation of a developmental sequence, as we will see. Before describing the regularities and variations observed in this set of simulations, let us first analyse a particular one where the developmental sequence is clearly observable. **Figure 7** exhibits such a simulation. We observe three clear developmental stages, i.e., three relatively homogeneous phases with rather sharp transitions. These stages are not pre-programmed, but emerge from the interaction of the vocal productions of the sensorimotor system, learning within the sensorimotor model, and the active choice of goals by intrinsically motivated active exploration. First (until $\approx 30,000$ vocalizations), the agent produces mainly motor commands which results in *no phonation* or in *unarticulated* vocalizations (in the sense of the classes defined section 2.1.3). Second (until $\approx 150,000$ vocalizations), phonation almost always occurs, but the vocalizations are mostly *unarticulated*. Third, it produces mainly *articulated* vocalizations.

The visualization of the developmental sequence of the 9 independent simulations, provided **Figure A2** in the Appendix, shows important interindividual variations whereas initial conditions are statistically similar due to initialization in line 1 of **Algorithm 1**. These variations can be understood through the

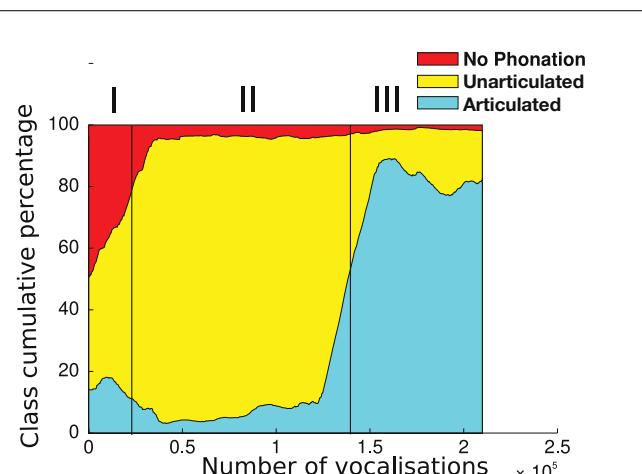


FIGURE 7 | Self-organization of vocal developmental stages. At each time step t (x-axis), the percentage of each vocalization class between t and $t + 30,000$ is plotted (y-axis), in a cumulative manner (sum to 100%). Vocalization classes are defined in section 2.1.3. Roman numerals show three distinct developmental stages. I: mainly no phonation or unarticulated vocalizations. II: mainly unarticulated. III: mainly articulated. The boundaries between these stages are not preprogrammed and are here manually set by the authors, looking at sharp transitions between relatively homogeneous phases.

⁴Each simulation involves several hours of computing on a desktop computer, due to the complexity of **Algorithm 1**, in particular in the Bayesian inference and update procedures.

interaction of the sensorimotor system f , the internal sensorimotor model G_{SM} and the interest model G_{IM} , resulting in a complex dynamical system where observed developmental sequences are particular attractors (see e.g., Van Geert, 1991; Smith and Thelen, 2003). Moreover the sensorimotor and the interest models are probabilistic, thus inducing a non-negligible source of variability all along a particular simulation. Another factor is that using an online learning process on a GMM can result in a sort of forgetting, leading sometimes to the re-exploration of previously learnt parts of the sensorimotor space⁵. However, the sequence *No phonation* → *Unarticulated* → *Articulated* appears as a global tendency, as shown in **Table 1**. We observe that despite variations, most simulations begin with a mix of *no phonation* and *unarticulated* vocalizations, then mainly produce *unarticulated* vocalizations, and often end up with *articulated* vocalizations. An analogy can be made with human phonological systems, which are all different in the details but display strong statistical tendencies (Maddieson and Precoda, 1989; Schwartz et al., 1997; Oudeyer, 2005; Moulin-Frier et al., 2011).

This suggests that the agent explores its sensorimotor space by producing vocalizations of increasing complexity. The class *no phonation* is indeed the easiest to learn to produce for two reasons: the rest positions of the pressure and voicing motor parameters do not allow phonation (both around -0.25 at the initialization of the agent, line 1 of **Algorithm 1**) ; and there is no variations on the formant values, which makes the control task trivial as soon

Table 1 | Count of vocalization stages in the 9 simulations of the supplementary data.

Types of sounds produced	Stage I	Stage II	Stage III	Stage IV
No phonation-unarticulated	7	0	2	0
Unarticulated	0	7	0	3
Articulated	0	2	4	0
Other	2	0	1	0

The “types of sounds produced” (first column of the table) correspond to the most prominent class in a given stage, where stages are manually set, looking at sharp transitions between relatively homogeneous phases. These developmental stages are therefore subjective to a certain extent, in the sense that another observer could have set different ones (but hopefully also would observe major structural changes). “No phonation-Unarticulated” means a mix between No phonation and Unarticulated classes (as defined in section 2.1.3 in that stage). A number x in a cell means this type of vocalizations (row) appears x times at the n^{th} stage of development (column) in the set of 9 simulations. Two to four developmental stages were identified in each simulation, explaining why the “Stage I” and “Stage II” columns sum up to 9 (the total number of simulations), but not the “Stage III” and “Stage IV” columns.

The bold number indicates the sequence (*No phonation* - *unarticulated*) → *Unarticulated* → *Articulated* is relatively stable across simulations.

⁵This is why we limited the simulations to 240,000 vocalizations each, in order to avoid this unwanted effect of forgetting. However, the fact that the system is able to adaptively re-explore sensorimotor regions that have been forgotten is an interesting feature of curiosity-driven learning.

as the agent has a bit of experience. There is more to learn with *unarticulated* vocalizations, where formant values are varying in at least one part of the vocalization, and still more with *articulated* ones where they are varying in both parts (for the first and second command).

Figure 8 shows what happens in the particular simulation of **Figure 7** in more details.

This developmental sequence is divided into 3 stages, I, II, and III, stages being separated by vertical dark lines on **Figure 8**, identical on each subplot (stage boundaries are the same than in **Figure 7**).

In stage I, until approximately 30,000 vocalizations, the agent produces mainly *no phonation* and *unarticulated* vocalizations. We observe that the agent set goals for $I_{(1)}$ either around 0, either around 1, whereas the goals for $I_{(2)}$ stay around 0 (last row in “Goals”). By trying to achieve these goals, the agent progressively refines its sensorimotor model and progresses by raising the values of the pressure and voicing motor parameter in the first command (two last rows of the section “Motor commands,” 1st column). Other articulators remain around the neutral position (value 0). The agent is learning to phonate. The percentages of vocalization belonging to each vocalization class is provided **Table 2**.

Then, in stage II, from 30,000 to approximately 150,000 vocalizations, the agent is mainly interested in producing vocalizations which begin with a *Vowels* [$I_{(1)} > 0.9$, see the definition of phone types in section 2.1.3] and finish with a *None* [$I_{(2)} < 0.1$]. An example of such a VN vocalization can be observed in the Appendix, **Figure A1** in section . During this stage, it learns to produce relatively high $F1_{(1)}$ values, in particular by decreasing the $Art_{1(1)}$ parameter (approximately controlling the jaw height, see **Figure 3**). Regarding the second command, although the agent self-generates various goals for $F1_{(2)}$ and $F2_{(2)}$, and produces various motor commands to try to reach them, the sound produced mostly corresponds to a *None* [$I_{(2)} = 0$, and therefore $F1_{(2)} = F2_{(2)} = 0$]. This is due both to the negative value of the voicing parameter (last row in “Motor commands,” second column), and to the fact that the vocal tract often ends in a closed configuration due to the poor quality of the sensorimotor model in this region (because phonation occurs very rarely for the second command, leaving the agent without an adequate learning set). During this stage, the agent explores a limited part of the sensorimotor space both in time (sound only for the first command) and space (around the neutral position), until it finally manages to phonate more globally at the end of this stage. This could be correlated to the acquisition of articulated vocalizations. The percentages of vocalization belonging to each vocalization class is provided in **Table 3**.

Finally, in stage III (until 150,000 to the end), phonation almost always occurs during both the perception time windows (see I densities, both for goals and reached values). An example of such a VV vocalization can be observed in the Appendix, **Figure A1** in section . This is much harder to achieve for two reasons: firstly because there is a need to control a sequence of 2 articulators movement in order to reach two formant values in sequence [i.e., $F1_{(1)}, F1_{(2)}, F2_{(1)}, F2_{(2)}$] instead of one in the previous stage (the second command leading to no sound), and

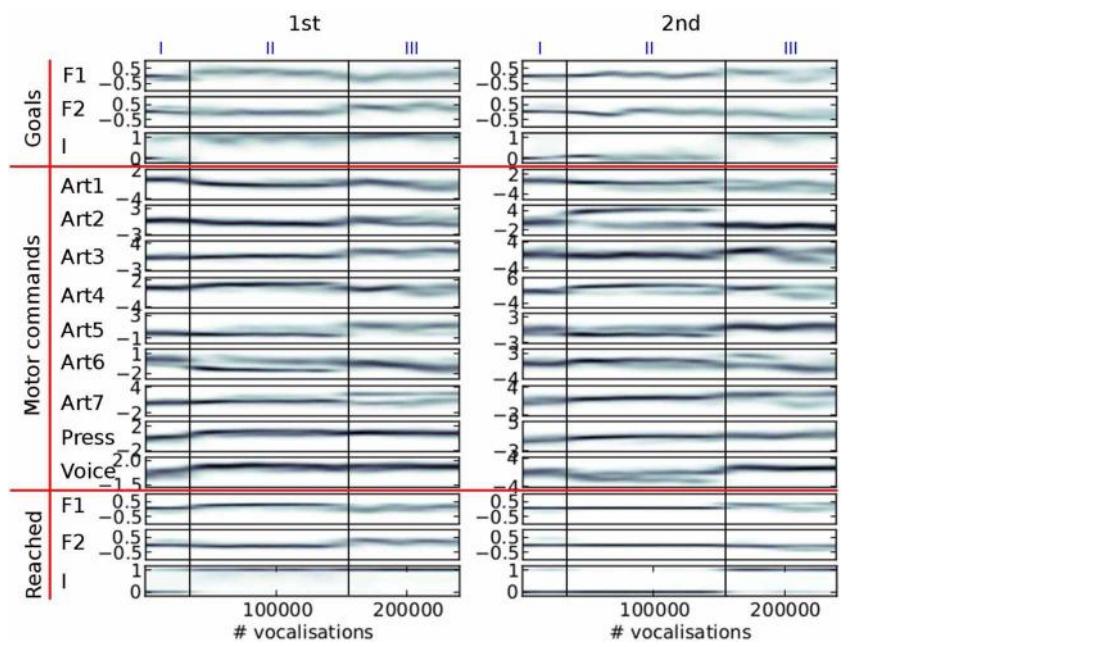


FIGURE 8 | Evolution of the distribution of auditory goals, motor commands and sounds actually produced over the life time of a vocal agent (the same agent as in Figure 7). The variables are in three groups (horizontal red lines): the goals chosen by the agent in line 3 of **Algorithm 1** (**top** group), the motor commands it inferred to reach the goals using its inverse model in line 4 (**middle** group), and the actual perceptions resulting from the motor commands through the synthesizer in line 5 (**bottom** group). There are two columns (1st and 2nd), because of the sequential nature of vocalizations (two motor commands per

vocalization). Each subplot shows the density of the values taken by each parameter (y-axis) over the life time of the agent (x-axis, in number of vocalizations since the start). It is computed using an histogram on the data (with 100 bins per axis), on which we apply a 3-bins wide Gaussian filter. The darker the color, the denser the data: e.g., the auditory parameter $I_{(2)}$ actually reached by the second command ($I_{(2)}$, last row in “Reached,” 2nd column), especially takes values around 0 (y-axis) until approximately 150,000th vocalization (x-axis), then it takes rather values around 1. The three developmental stages of **Figure 7** are reported at the **top**.

secondly because the position of the articulators reached for the second command also depends on the position of the articulators reached for the first one (a kind of coarticulation due to the dynamical properties of the motor system). We observe that the range of values explored in the sensorimotor space is larger than for the previous stage (both in motor and auditory spaces). The percentages of vocalizations belonging to each vocalization class is provided in **Table 4**.

3.2. INFLUENCE OF THE AUDITORY ENVIRONMENT

In a second set of experiments, we integrated a social environment providing a set of adult vocalizations. As explained in section 2.4, the learner has an additional choice: it can explore autonomously, or emulate the adult vocalizations. An “ambient language” is here modeled as a set of two speech sounds. To make it coherent with human language and the learning process observed in development, we chose speech-like sounds, typically vowel or consonant-vowel sounds. In terms of our sensorimotor descriptions, the adult sounds correspond to I_1 with low values and I_2 with high values. **Figure 9** shows such vocalizations corresponding to those used by Teacher 1 in **Figure 10**.

Figure 10 shows a significant evolution in the agent’s vocalizations. In the early stage of its development, it can only make a few sounds. Most sounds correspond to small values of $I_{1(1)}, F_{1(1)}, F_{1(2)}, F_{2(1)}$, and $F_{2(2)}$, as in the first developmental

Table 2 | Percentage of vocalization classes produced in stage I of the studied developmental sequence.

NN	CN	NC	VN	NV	VV	CV	VC	CC
45.3%	13.4%	0.6%	18.9%	4.5%	9.9%	6.6%	0.7%	0.2%

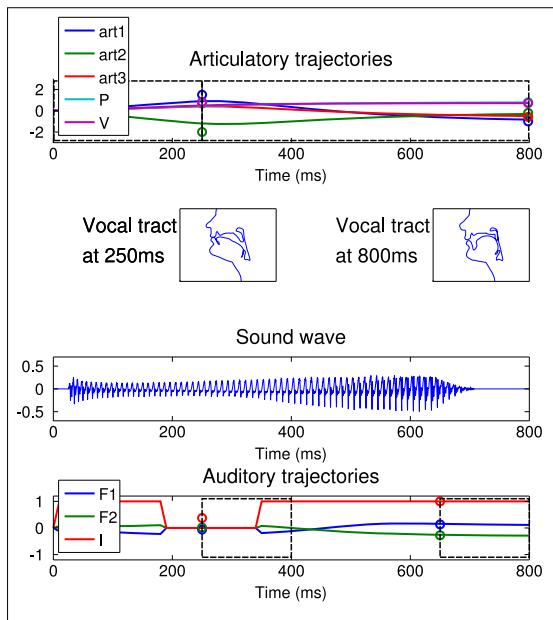
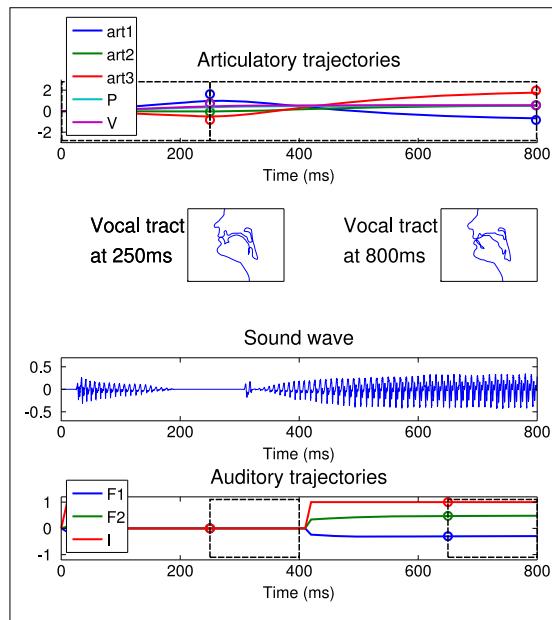
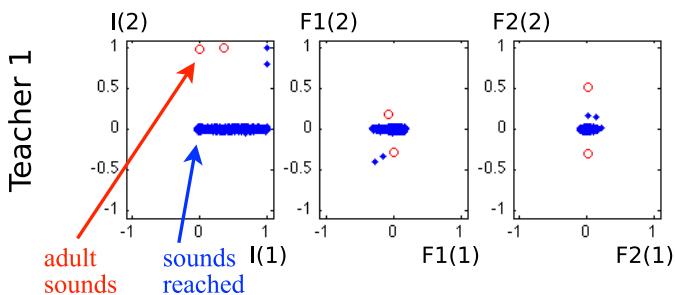
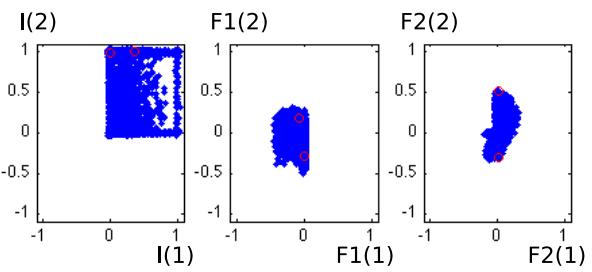
Table 3 | Percentage of vocalization classes produced in stage II of the studied developmental sequence.

NN	CN	NC	VN	NV	VV	CV	VC	CC
4.0 %	26.9 %	0.1 %	62.2 %	0.1 %	3.4 %	0.5 %	2.5 %	0.2 %

Table 4 | Percentage of vocalization classes produced in stage III of the studied developmental sequence.

NN	CN	NC	VN	NV	VV	CV	VC	CC
1.6 %	3.7 %	0.1 %	12.1 %	0.8 %	67.5 %	6.5 %	6.8 %	0.8 %

stage of the previous experiment (see **Table 2** and **Figure 8**). Therefore the agent is not able to reproduce the ambient sounds of its environment. In contrast, in later periods of its development, its vocalizations cover a wider range of sounds, with

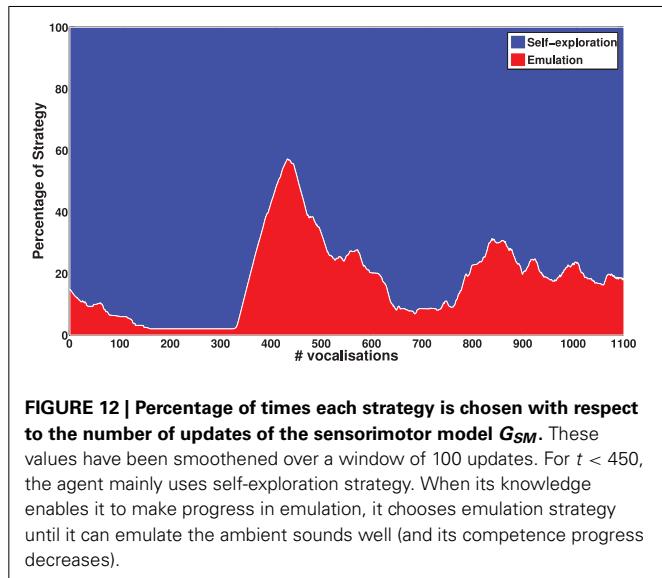
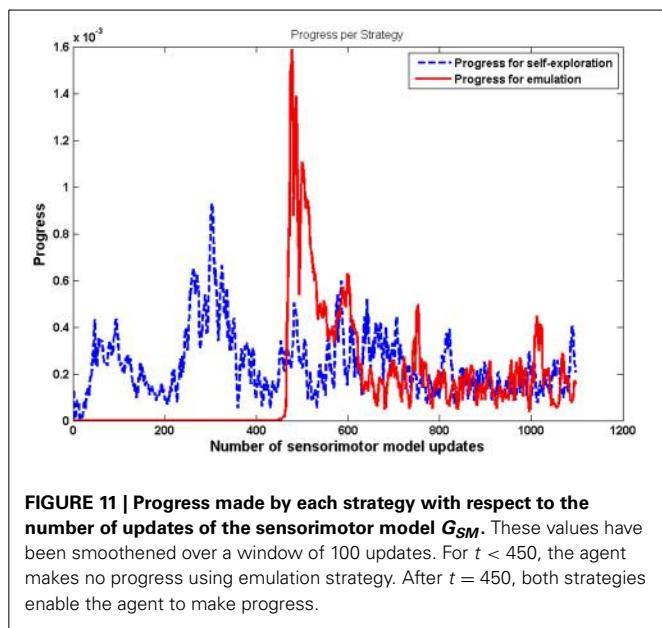
Adult vocalization 1**Adult vocalization 2****FIGURE 9 |** The two vocalizations of the adult Teacher 1 used in Figure 10, with the same convention as in Figure 4.**A First vocalizations****B Mature vocalizations****FIGURE 10 |** Vocalizations of the learning agent in the early and mature stages of vocal development. **(A)** All auditory outcomes produced by the agent in its early stage of vocalization are represented by blue dots in the 6-dimensional space of the auditory outcomes. The adult sounds are represented in red circles. The actually produced auditory outcomes only cover a small area of

physically possible auditory outcomes, and correspond mostly to $I_{(2)} = 0$, which represent vowel-consonant or consonant-consonant types of syllables. **(B)** The auditory outcomes produced by the infant in its mature stage of vocalization cover a much larger area of auditory outcomes and extend in particular over areas in which vocalizations of the social peer are located.

notably $I_{(1)}$ and $I_{(2)}$ both positive, which means it now produces more articulated sounds. The development of vocalizations for a self-exploring agent in the last section showed that it progressively was able to produce articulated vocalizations, which we observed at times at the end of its development. This effect has been reinforced by the environment: with articulated vocalizations to emulate, it produces this class more regularly.

Another important result is that mature vocalizations can now reproduce the ambient sounds of the environment: the regions of the sounds produced by the learner (blue dots) overlap the teacher's demonstrations (red circles). It seems that, during the

first vocalizations, the agent cannot emulate the ambient sounds because they are too far away from its possible productions, and thus it can hardly make any progress and approach these demonstrations. **Figure 11** confirms this interpretation. In the beginning, the agent makes no progress with emulation, and it is only around $t = 450$ that it makes progress with the emulation strategy. At that point, as we can see in **Figure 12**, it uses equally both strategies. This enables the agent to make considerable progress from $t = 450$ to $t = 800$. Indeed, once its mastery improves and the set of sounds it can produce increases, it then increasingly emulates ambient sounds. Once it manages to emulate the



ambient sounds well, and thus its competence progress decreases, it uses less the emulation strategy and more the self-exploration strategy.

To analyse better this emulation phenomenon and assess the influence of the ambient language, we run the same experiment with different acoustic environments. We used two other sets of speech sound demonstrations from simulated peers, and analysed the auditory productions of the agent in **Figure 13**. The first property that can be noted is that in the early phase of the vocal exploration (**Figures 13A,C**), the auditory productions of the two agents are alike, and do not depend on the speech environment. On the contrary, the mature vocalizations vary with respect to the speech environment. With Teacher 1, the productions have their values $F2_{(1)}$ and $F2_{(2)}$ along the axis formed by the demonstration (**Figure 10A**, last column). Comparatively, Teacher 2's speech

sounds have different $F1_{(1)}$, $F1_{(2)}$, $F2_{(1)}$, and $F2_{(2)}$. As represented in **Figure 13B**, the two speech sounds now differ mainly by their $F1_{(1)}$ (instead of $F1_{(2)}$) and in their subspace [$F2_{(1)}$, $F2_{(2)}$] the speech sounds have approximately rotated from those of Teacher 1. The produced auditory outcomes of the learner look like they have changed in the same way. Whereas the reached space (blue area) seemed to be along axis $F1_{(2)}$ and $F2_{(2)}$ and little on $F1_{(1)}$ or $F2_{(1)}$ for Teacher 1, it has extended its exploration along $F1_{(2)}$ and $F2_{(2)}$ for Teacher 2. With Teacher 3, the demonstrations are more localized in the auditory space, with $F1_{(1)} < 0$ and $F2_{(2)} > 0$. The effect we observe in **Figure 13D** is that the exploration is more localized too: the explored space is more oriented toward areas where $F1_{(1)} < 0$ and $F2_{(2)} > 0$. Thus, these three examples strongly suggest a progressive influence of the auditory environment, in the sense that the first vocalizations in **Figures 10, 13** are very similar, whereas we observe a clear influence of the speech environment on the produced vocalizations in later stages.

Altogether, the results of these experiments provide a computational support to the hypothesis that the progressive influence of the ambient language observed in infant vocalizations can be driven by an intrinsic motivation to maximize competence progress. At early developmental stages, attempts to imitate adult vocalizations are certainly largely unsuccessful because basic speech principles, such as phonation, are not yet mastered. In this case, focusing on simpler goals probably yields better progress niches than an imitative behavior. While they are progressively mastered, the interest in these goals decreases whereas the ability to imitate adult vocalizations increases. Imitation thus becomes a new progress niche to explore.

4. DISCUSSION

Our main contribution with respect to previous computational models of speech acquisition is that we do not presuppose the existence of successive developmental stages, but rather they can emerge from an intrinsic drive to maximize the competence progress. We showed that vocal developmental stages can self-organize autonomously, from simple sensorimotor activities to more complex ones. The agent starts producing *no phonation* and *unarticulated* vocalizations, which are easy to produce because limited in the range of their auditory effects. This can be related to the first stage in infant vocal development (**Figure 1**), where the agent produces non speech-sounds (e.g., growls, squeals...) before learning phonation and then produces not well-articulated quasi-vowels. Later on, once the agent does not progress much in producing *unarticulated* vocalizations, it focuses on more complex vocalizations of the *articulated* class. The reason is that, due to the properties of the sensorimotor system and internal model, the mastering of complex tasks require first the mastering of simpler tasks in order to yield significant competence progress, so that these complex tasks are selected as interesting goals.

We also showed that intrinsically motivated exploration can lead to a progressive interest toward the sounds of the ambient language. Whereas the first vocalizations are mainly the result of self-exploration, they progressively lead to mastering necessary speech principles (e.g., phonation). This progressive mastering allows in turn to make significant progress in adult-speech

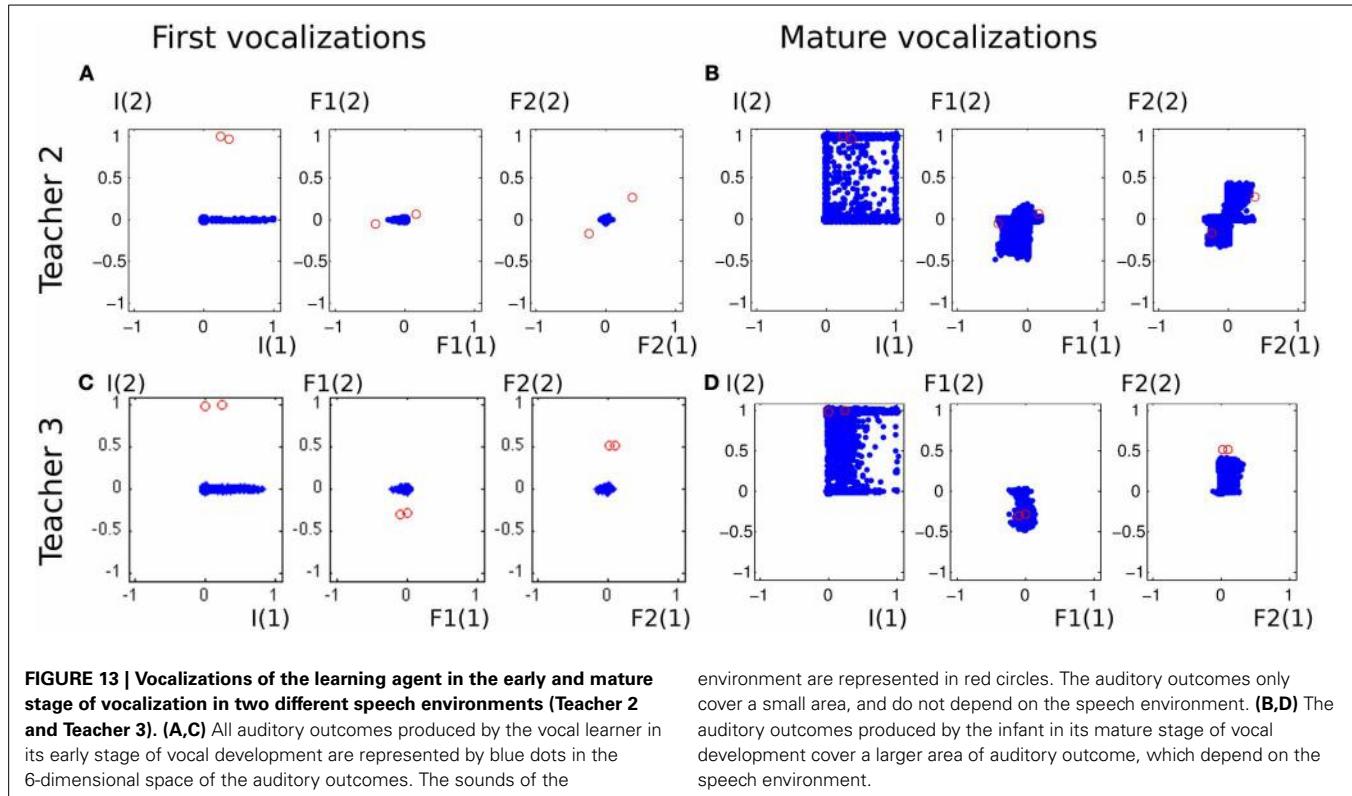


FIGURE 13 | Vocalizations of the learning agent in the early and mature stage of vocalization in two different speech environments (Teacher 2 and Teacher 3). (A,C) All auditory outcomes produced by the vocal learner in its early stage of vocal development are represented by blue dots in the 6-dimensional space of the auditory outcomes. The sounds of the

environment are represented in red circles. The auditory outcomes only cover a small area, and do not depend on the speech environment. (B,D) The auditory outcomes produced by the infant in its mature stage of vocal development cover a larger area of auditory outcome, which depend on the speech environment.

imitation, which explains why the vocal learner starts to choose more often as targets the sound of its environment. Competence-progress based curiosity-driven exploration could thus explain a progressive influence of the ambient language on infant vocalizations.

We therefore showed that intrinsically motivated active exploration can self-organize a coherent developmental sequence, without any external clock or preset specification of this sequence. This possible role of intrinsic motivation, providing a mechanism to discover autonomously necessary developmental stages to structure the learning process, is here validated computationally. We believe that it could be of major interest for understanding the structuration of early vocal development in infants. Speech acquisition is such a complex task that intrinsic motivation could be a crucial component to make it possible in the infant's first year of life.

Our model, however, has a number of limitations. Firstly, our modeling choices of the articulatory and auditory representations, as well as the implementation of the transformation from the former to the latter, is somewhat less realistic than in some previous models: articulatory trajectories are specified using two commands per articulator with fixed durations and the auditory representation uses only three acoustic parameters (the intensity and the two first formants) averaged in fixed and relatively arbitrary perception time windows. Moreover, the fact that formant values are set to 0 whenever the intensity of the signal is null can appear quite unrealistic. Although previous models often provide more meticulous implementations of the sensorimotor system, including e.g., pitch or tactile information, what is important

to us is a sensorimotor system where all vocalizations are not equally easy to learn in terms of control. Such a requirement is certainly necessary for a clear developmental sequence to emerge. Secondly, we did not treat a major issue in speech acquisition research, the so-called correspondence problem: how the child is able to relate its own vocalizations to adult vocalizations, whereas the vocal tract of the child is very different in size and geometry than the one of an adult, and therefore the spectral characteristics of the produced sounds are different. Solutions to overcome this problem have been proposed, generally based on adult feedback or reformulations associated with infant productions (Ishihara et al., 2009; Howard and Messum, 2011; Miura et al., 2012). This is outside the scope of this paper where our focus is on the self-organization of the developmental sequence in successive stages of increasing complexity. Extending our model to the interaction with real humans would definitely require to consider this issue.

Further works will consider higher-dimensional sensorimotor spaces for more realism. For example, the free software Praat (Boersma, 2012) is a powerful tool allowing to synthesize a speech signal from a trajectory in a 29-dimensional space of respiratory and oro-facial muscles. Numerous acoustic features can in turn be extracted from the synthesized sound, among which the Mel-frequency cepstral coefficients (MFCC; Davis and Mermelstein, 1980). It would also be interesting to study the effect of a vocal tract growing during the learning process, to study if our intrinsically motivated agent could re-explore only parts of the sensorimotor space which were the most affected by the vocal tract shape change. Generally, we believe that a developmental robotics approach applied to a realistic articulatory model can

appropriately manage the learning process of a complex and changing mapping in high-dimensional spaces, and that observed developmental sequences can lead to interesting comparisons with infant data and predictions. Regarding the present study, such a prediction could be that a human infant should be influenced by adult sounds earlier if they were easier to produce than well-formed syllables. For example, one could imagine an experiment in which a very young infant is put in an environment where he hears external sounds that are simpler than vowels/consonants/syllables (e.g., growls) and test whether his vocalizations become influenced by external environment earlier and/or if we can measure a greater interest than in a normal speech environment.

ACKNOWLEDGMENTS

The authors would like to thank Louis-Jean Boë for the design of **Figure 2** (vocal tract by Sophie Jacopin).

FUNDING

This work was partially financed by ERC Starting Grant EXPLORERS 240 007.

REFERENCES

- Baldassarre, G. (2011). "What are intrinsic motivations? a biological perspective," in *IEEE International Conference on Development and Learning (ICDL)*, Vol. 2, 1–8. doi: 10.1109/DEVLRN.2011.6037367
- Baldassarre, G., and Mirolli, M. (2013). *Intrinsically Motivated Learning in Natural and Artificial Systems*. Berlin; Heidelberg: Springer. doi: 10.1007/978-3-642-32375-1
- Baranes, A., and Oudeyer, P.-Y. (2010). "Intrinsically motivated goal exploration for active motor learning in robots: a case study," in *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2010)* (Taipei).
- Baranes, A., and Oudeyer, P.-Y. (2013). Active learning of inverse models with intrinsically motivated goal exploration in robots. *Robot. Auton. Syst.* 61, 49–73. doi: 10.1016/j.robot.2012.05.008
- Barto, A., Singh, S., and Chenetaz, N. (2004). "Intrinsically motivated learning of hierarchical collections of skills," in *Proc. 3rd Int. Conf. Dvp. Learn.*, San Diego, CA. 112–119.
- Berlyne, D. E. (1954). A theory of human curiosity. *Br. J. Psychol.* 45, 180–191.
- Boersma, P. (1998). *Functional Phonology: Formalizing the Interactions Between Articulatory and Perceptual Drives*. The Hague: Holland Academic Graphics.
- Boersma, P., and Weenink, D. . (2012). Praat: doing phonetics by computer [computer program]. Available online at: <http://www.praat.org/>
- Calinon, S. (2009). *Robot Programming by Demonstration*. CRC. Available online at: http://www.ppur.org/produit/505/9782940222315/Robot_20_Programming_20_by_20_Demonstration
- Call, J., and Carpenter, M. (2002). "Imitation in animals and artifacts," in *Chapter Three Sources of Information in Social learning*, eds C. L. Nehaniv and K. Dautenhahn (Cambridge, MA: MIT Press), 211–228.
- Csikszentmihalyi, M. (1997). *Creativity: Flow and the Psychology of Discovery and Invention*. New York, NY: HarperCollins.
- Davis, S., and Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. Acoust. Speech Signal Process.* 28, 357–366. doi: 10.1109/TASSP.1980.1163420
- Deci, E., and Ryan, R. M. (1985). *Intrinsic Motivation and Self-Determination in Human Behavior*. New York, NY: Plenum Press. doi: 10.1007/978-1-4899-2271-7
- Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. B (Methodol.)* 39, 1–38.
- Friston, K., Adams, R. A., Perrinet, L., and Breakspear, M. (2012). Perceptions as hypotheses: saccades as experiments. *Front. Psychol.* 3:151. doi: 10.3389/fpsyg.2012.00151
- Gottlieb, G. (1991). Experiential canalization of behavioral development: theory. *Dev. Psychol.* 27, 4. doi: 10.1037/0012-1649.27.1.4
- Gottlieb, J., Oudeyer, P.-Y., Lopes, M., and Baranes, A. (2013). Information-seeking, curiosity, and attention: computational and neural mechanisms. *Trends Cogn. Sci.* 17, 585–593. doi: 10.1016/j.tics.2013.09.001
- Guenther, F. H. (2006). Cortical interactions underlying the production of speech sounds. *J. Commun. Disord.* 39, 350–365. doi: 10.1016/j.jcomdis.2006.06.013
- Guenther, F. H., Ghosh, S. S., and Tourville, J. A. (2006). Neural modeling and imaging of the cortical interactions underlying syllable production. *Brain Lang.* 96, 280–301. doi: 10.1016/j.bandl.2005.06.001
- Guenther, F. H., Hampson, M., and Johnson, D. (1998). A theoretical investigation of reference frames for the planning of speech movements. *Psychol. Rev.* 105, 611–633. doi: 10.1037/0033-295X.105.4.611-633
- Hart, S. (2009). "An intrinsic reward for affordance exploration," in *ICDL International Conference on Developmental Learning*, (Shanghai).
- Howard, I., and Messum, P. (2011). Modeling the development of pronunciation in infant speech acquisition. *Motor Control* 15, 85–117. Available online at: <http://journals.human kinetics.com/mc-back-issues/mc-volume-15-issue-1-january/modeling-the-development-of-pronunciation-in-infant-speech-acquisition>
- Ishihara, H., Yoshikawa, Y., Miura, K., and Asada, M. (2009). How caregiver's anticipation shapes infant's vowel through mutual imitation. *IEEE Trans. Auton. Ment. Dev.* 1, 217–225. doi: 10.1109/TAMD.2009.2038988
- Kaplan, F., and Oudeyer, P.-Y. (2007a). "The progress-drive hypothesis: an interpretation of early imitation," in *Models and Mechanisms of Imitation and Social Learning: Behavioural, Social and Communication Dimensions*, eds K. Dautenhahn, and C. Nehaniv (Cambridge: Cambridge University Press), 361–378. doi: 10.1017/CBO9780511489808.024
- Kaplan, F., and Oudeyer, P.-Y. (2007b). In search of the neural circuits of intrinsic motivation. *Front. Neurosci.* 1:1. doi: 10.3389/neuro.01.1.1.017.2007
- Kröger, B. J., Kannampuzha, J., and Neuschafer-Rube, C. (2009). Towards a neurocomputational model of speech production and perception. *Speech Commun.* 51, 793–809. doi: 10.1016/j.specom.2008.08.002
- Kuhl, P. K. (2004). Early language acquisition: cracking the speech code. *Nat. Rev. Neurosci.* 5, 831–843. doi: 10.1038/nrn1533
- Lopes, M., Melo, F., Montesano, L., and Santos-Victor, J. (2010). "Abstraction Levels for Robotic Imitation: Overview and Computational Approaches," in *From Motor Learning to Interaction Learning in Robots*. Vol. 264, eds O. Sigaud and J. Peters (Berlin; Heidelberg: Springer), 313–355. doi: 10.1007/978-3-642-05181-4_14
- Lopes, M., and Oudeyer, P.-Y. (2012). "The strategic student approach for life-long exploration and learning," in *IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL)*, (San Diego, CA), 1–8. doi: 10.1109/DevLrn.2012.6400807
- Maddieson, I., and Precoda, K. (1989). Updating UPSID. *J. Acoust. Soc. Am.* 86, S19. doi: 10.1121/1.2027403
- Maeda, S. (1989). Compensatory articulation during speech: evidence from the analysis and synthesis of vocal tract shapes using an articulatory model. *Speech Prod. Speech Model* 55, 131–149.
- Markey, K. L. (1994). *The Sensorimotor Foundations Of Phonology: A Computational Model of Early Childhood Articulatory and Phonetic Development*. PhD thesis, University of Colorado at Boulder.
- Merrick, K., and Maher, M. L. (2009). Motivated learning from interesting events: adaptive, multitask learning agents for complex environments. *Adapt. Behav.* 17, 7–27. doi: 10.1177/1059712308100236
- Miura, K., Yoshikawa, Y., and Asada, M. (2012). Vowel acquisition based on an auto-mirroring bias with a less imitative caregiver. *Adv. Robot.* 26, 23–44. doi: 10.1163/016918611X607347
- Moulin-Frier, C., and Oudeyer, P.-Y. (2012). "Curiosity-driven phonetic learning," in *International Conference on Development and Learning, Epirob* (San Diego, CA).
- Moulin-Frier, C., and Oudeyer, P.-Y. (2013a). "Exploration strategies in developmental robotics: a unified probabilistic framework," in *International Conference on Development and Learning, Epirob*, Osaka.
- Moulin-Frier, C., and Oudeyer, P.-Y. (2013b). "The role of intrinsic motivations in learning sensorimotor vocal mappings: a developmental robotics study," in *Proceedings of Interspeech*, (Lyon).
- Moulin-Frier, C., Schwartz, J., Diard, J., and Bessière, P. (2011). "Emergence of articulatory-acoustic systems from deictic interaction games in a "Vocalize to Localize" framework," in *Primate Communication and Human Language*:

- Vocalisations, Gestures, Imitation and Deixis in Humans and Non-humans, Advances in Interaction studies*, eds A. Vilain, J.-L. Schwartz, C. Abry, and J. Vauclair (Amsterdam: John Benjamins Pub. Co.), 193–220. Available online at: <https://benjamins.com/#catalog/books/ais.1/main>
- Nehaniv, C. L., and Dautenhahn, K. (2007). *Imitation and Social Learning in Robots, Humans and Animals: Behavioural, Social and Communicative Dimensions*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511489808
- Nguyen, S. M., and Oudeyer, P.-Y. (2012). Active choice of teachers, learning strategies and goals for a socially guided intrinsic motivation learner. *Paladyn J. Behav. Robot.* 3, 136–146. doi: 10.2478/s13230-013-0110-z
- Oller, D. K. (2000). *The Emergence of the Speech Capacity*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Oudeyer, P. (2005). The self-organization of speech sounds. *J. Theor. Biol.* 233, 435–449. doi: 10.1016/j.jtbi.2004.10.025
- Oudeyer, P.-Y., Baranes, A., Kaplan, F., and Ly, O. (2013). *Intrinsically Motivated Learning in Natural and Artificial Systems, Chapter Developmental Constraints on Intrinsically Motivated Skill Learning: Towards Addressing High-Dimensions and Unboundedness in the Real World*. Springer.
- Oudeyer, P.-Y., and Kaplan, F. (2006). Discovering communication. *Connect. Sci.* 18, 189–206. doi: 10.1080/09540090600768567
- Oudeyer, P.-Y., and Kaplan, F. (2007). What is intrinsic motivation? a typology of computational approaches. *Front. Neurorobotics* 1:6. doi: 10.3389/neuro.12.006.2007
- Oudeyer, P.-Y., Kaplan, F., and Hafner, V. (2007). Intrinsic motivation systems for autonomous mental development. *IEEE Trans. Evol. Comput.* 11, 265–286. doi: 10.1109/TEVC.2006.890271
- Ryan, R. M., and Deci, E. L. (2000). Intrinsic and extrinsic motivations: classic definitions and new directions. *Contemp. Educ. Psychol.* 25, 54–67. doi: 10.1006/ceps.1999.1020
- Schembri, M., Mirolli, M., and Baldassarre, G. (2007). “Evolving childhoods length and learning parameters in an intrinsically motivated reinforcement learning robot,” in *Proceedings of the Seventh International Conference on Epigenetic Robotics*, Vol. 134 (Lund: Lund University), 141–148.
- Schmidhuber, J. (1991). “A possibility for implementing curiosity and boredom in model-building neural controllers,” in *Proc. SAB’91*, eds J. A. Meyer and S. W. Wilson (Cambridge: MIT Press), 222–227.
- Schmidhuber, J. (2010). Formal theory of creativity, fun, and intrinsic motivation (1990–2010). *IEEE Trans. Auton. Ment. Dev.* 2, 230–247. doi: 10.1109/TAMD.2010.2056368
- Schwartz, J.-L., Boë, L.-J., Vallée, N., and Abry, C. (1997). Major trends in vowel system inventories. *J. Phon.* 25, 233–253. doi: 10.1006/jpho.1997.0044
- Sigismund, B. (1971). *Child Language: A Book of Readings, Chapter Kind und Welt*. Englewood Cliffs, NJ: Prentice-Hall. (Original work published in 1856).
- Smith, L. B., and Thelen, E. (2003). Development as a dynamic system. *Trends Cogn. Sci.* 7, 343–348. doi: 10.1016/S1364-6613(03)00156-6
- Srivastava, R. K., Steunebrink, B. R., and Schmidhuber, J. (2013). First experiments with powerplay. *Neural Netw.* 41, 130–136. doi: 10.1016/j.neunet.2013.01.022
- Stout, A., and Barto, A. G. (2010). “Competence progress intrinsic motivation,” in *IEEE 9th International Conference on Development and Learning (ICDL)*, (Ann Arbor), 257–262.
- Taine, H. (1971). *Child Language: A Book of Readings, Chapter Acquisition of Language by Children*. Englewood Cliffs, NJ: Prentice-Hall. (Original work published in 1856).
- Thelen, E., and Smith, L. (1996). *A Dynamic Systems Approach to the Development of Cognition and Action*. Cambridge: A Bradford book; MIT Press. Available online at: <http://mitpress.mit.edu/books/dynamic-systems-approach-development-cognition-and-action>
- Van Geert, P. (1991). A dynamic systems model of cognitive and language growth. *Psychol. Rev.* 98, 3. doi: 10.1037/0033-295X.98.1.3
- Vihman, M. M., Ferguson, C. A., and Elbert, M. (1986). Phonological development from babbling to speech: common tendencies and individual differences. *Appl. Psycholinguist.* 7, 3–40. doi: 10.1017/S0142716400007165
- Warlaumont, A. (2012). “A spiking neural network model of canonical babbling development,” in *IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL)*, (San Diego, CA), 1–6.
- Warlaumont, A. (2013a). “Salience-based reinforcement of a spiking neural network leads to increased syllable production,” in *IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL)*, (Osaka), 1–7. doi: 10.1109/DevLrn.2013.6652547
- Warlaumont, A. S. (2013b). Prespeech motor learning in a neural network using reinforcement. *Neural Netw.* 38, 64–95.
- Whiten, A. (2000). Primate culture and social learning. *Cogn. Sci.* 24, 477–508. doi: 10.1207/s15516709cog2403_6

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 03 July 2013; accepted: 17 December 2013; published online: 16 January 2014.

Citation: Moulin-Frier C, Nguyen SM and Oudeyer P-Y (2014) Self-organization of early vocal development in infants and machines: the role of intrinsic motivation. *Front. Psychol.* 4:1006. doi: 10.3389/fpsyg.2013.01006

This article was submitted to Cognitive Science, a section of the journal *Frontiers in Psychology*.

Copyright © 2014 Moulin-Frier, Nguyen and Oudeyer. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

APPENDIX

VOCALIZATION TYPES

Figure A1 shows the 9 types of vocalizations defined in section 2.1.3 (NN, CN, NC, VN, NV, VV, VC, CV and CC).

DEVELOPMENTAL SEQUENCES OF 9 INDEPENDENT SIMULATIONS

The figures of this section display the emerging developmental sequence of 9 independent simulations in pure self-exploration mode (section 3.1). At each time step t (x-axis), the percentage of

each vocalization class during between t and $t + 30,000$ is plotted (y-axis), in a cumulative manner. Vocalization classes are defined in section 2.1.3. For each one, we show boundaries between developmental stages. These boundaries are set manually, by looking at sharp transitions between relatively homogeneous phases. They are therefore subjective to a certain extent, in the sense that another observer could have set different ones (but hopefully also would observe major structural changes).

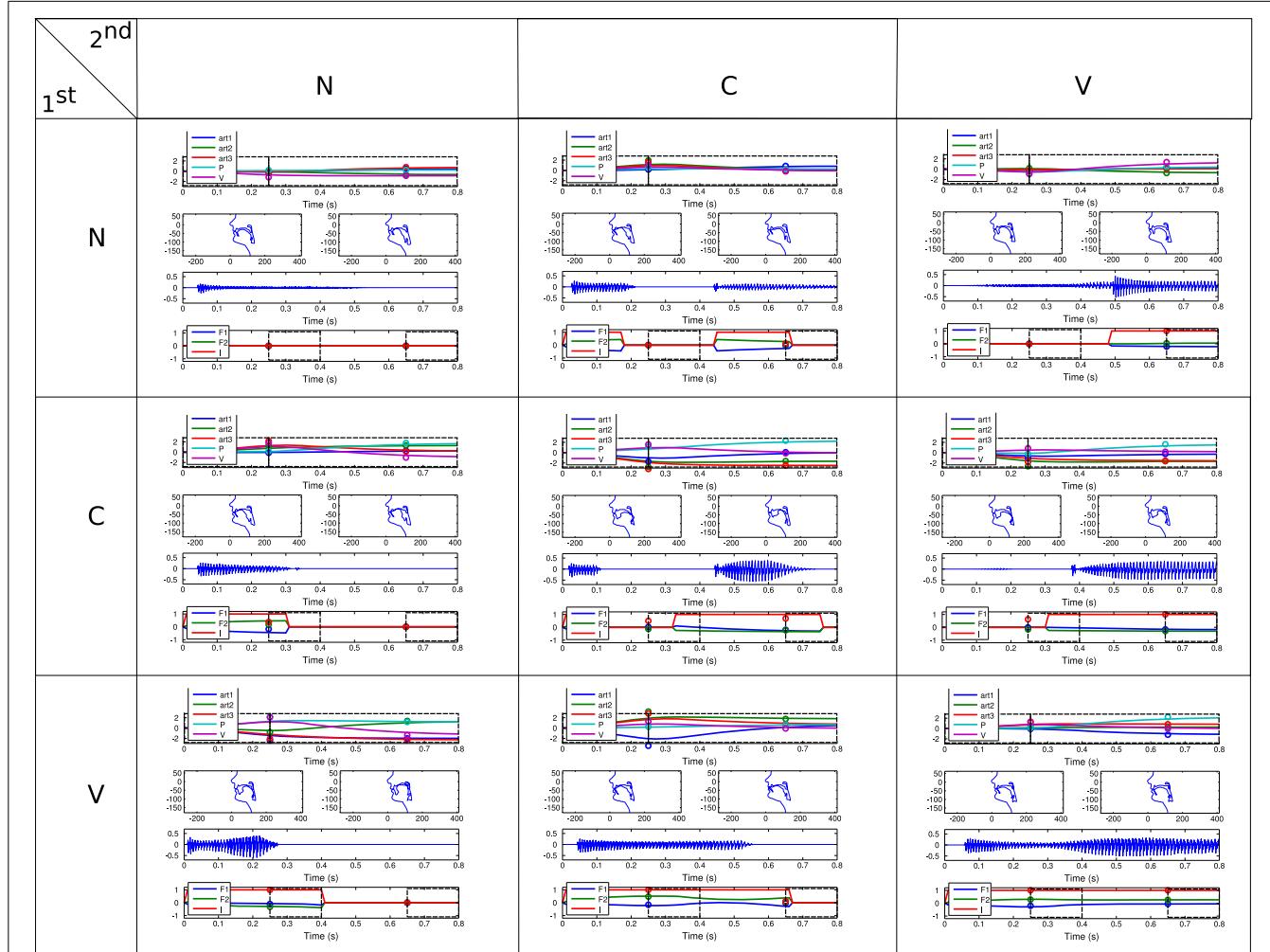


FIGURE A1 | Examples of each vocalization types. Rows (1st) correspond to the type of the first phone and columns (2nd) to the type of the second phone of the vocalization. There are three possible phone types, as defined in section 2.1.3: the *Vowels* (V) which have a high

intensity ($I > 0.9$), the *Consonants* (C) which have a low intensity ($0.1 < I < 0.9$) and the *None* which have almost no intensity ($I < 0.1$). For example, the plot in the second row (C) third column (V) corresponds to a CV vocalization, with the same convention as in **Figure 4**.

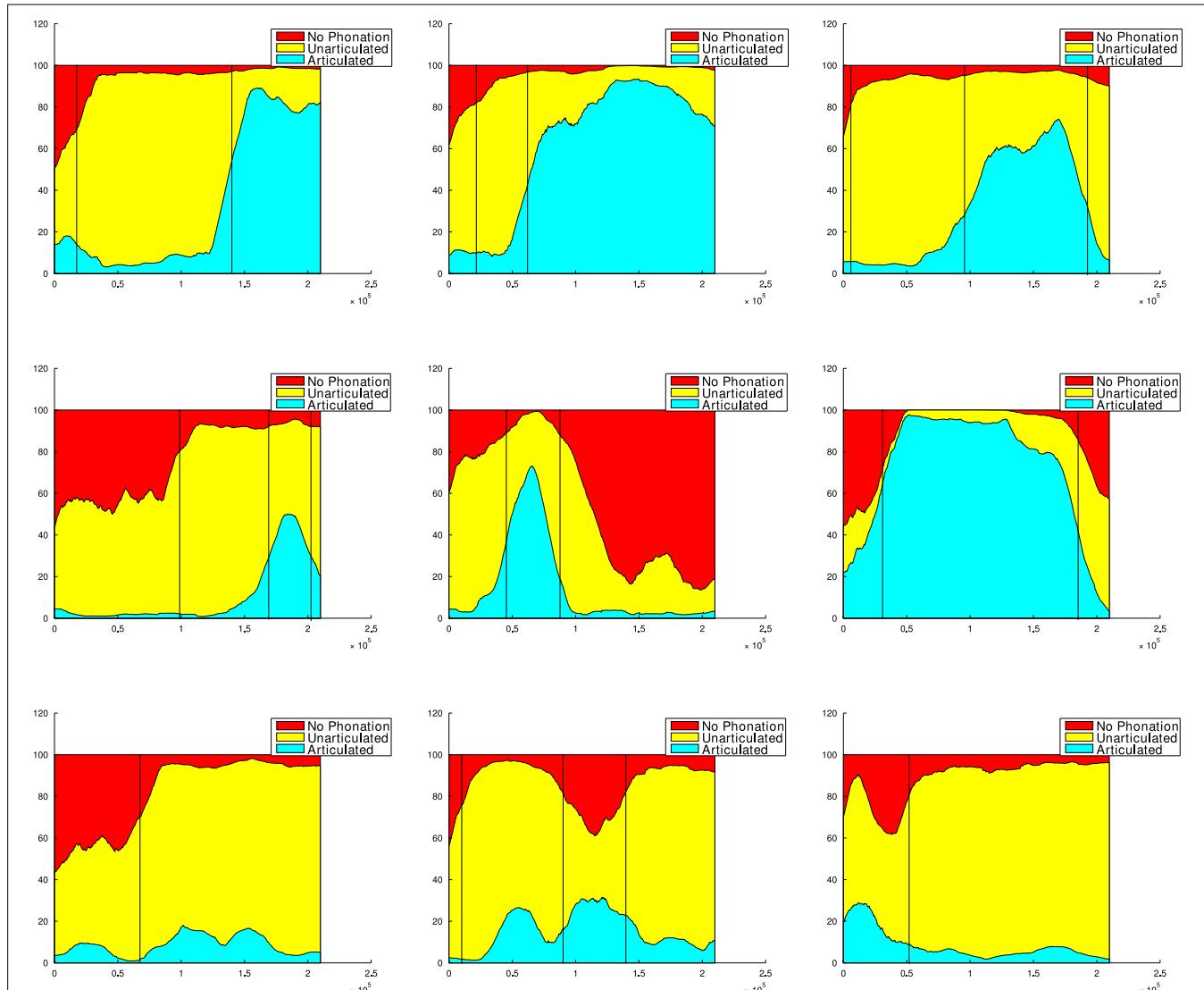


FIGURE A2 | Developmental sequences emerging from the 9 simulations for the experiment described in section 3.1.
Each subplot follows the same convention as in **Figure 7**.
The simulations have been ordered, also in a subjective

manner, from those which display a clear developmental sequence of the type *No phonation* → *Unarticulated* → *Articulated* to those less organized (from **left** to **right**, then **top** to **bottom**).



A motivation model for interaction between parent and child based on the need for relatedness

Masaki Ogino^{1*}, Akihiko Nishikawa² and Minoru Asada²

¹ Department of Informatics, Kansai University, Osaka, Japan

² Department of Adaptive Machine Systems, Graduate School of Engineering, Osaka University, Osaka, Japan

Edited by:

Richard M. Ryan, University of Rochester, USA

Reviewed by:

Tom Stafford, University of Sheffield, UK

Sufen Chen, Albert Einstein College of Medicine, USA

***Correspondence:**

Masaki Ogino, Department of Informatics, Kansai University, 2-1-1 Ryozenjicho, Takatsuki, Osaka 569-1095, Japan

e-mail: ogino@kansai-u.ac.jp

In parent-child communication, emotions are evoked by various types of intrinsic and extrinsic motivation. Those emotions encourage actions that promote more interactions. We present a motivation model of infant-caregiver interactions, in which relatedness, one of the most important basic psychological needs, is a variable that increases with experiences of emotion sharing. Besides being an important factor of pleasure, relatedness is a meta-factor that affects other factors such as stress and emotional mirroring. The proposed model is implemented in an artificial agent equipped with a system to recognize gestures and facial expressions. The baby-like agent successfully interacts with an actual human and adversely reacts when the caregiver suddenly ceases facial expressions, similar to the "still-face paradigm" demonstrated by infants in psychological experiments. In the simulation experiment, two agents, each controlled by the proposed motivation model, show relatedness-dependent emotional communication that mimics actual human communication.

Keywords: intrinsic motivation, relatedness, interaction, emotion

1. INTRODUCTION

Humans acquire knowledge and skills voluntarily by interacting with the environment. This voluntary learning process is driven by intrinsic motivation, which embodies curiosity and interest. By contrast, extrinsic motivation results in rewards such as food. White (1959) proposed that the intrinsic desire to interact with the environment and others underlies human exploratory behavior. Intrinsic motivation encourages individuals to seek novelty, uncertainty, and complexity (Berlyne, 1960). According to the self-determination theory of Ryan and Deci (2000), humans have three inherent fundamental needs: autonomy, competence, and relatedness. Autonomy is the perception that one's behavior is compatible with one's approval. Competence is fulfilled when expected or desired results are achieved. Relatedness is gained when one senses a close relationship with others. Ryan and Deci insist that these fundamental needs and individual differences are shaped by the social context.

Fundamental needs are closely related to emotions. Reis et al. (2000) showed that satisfaction levels of fundamental needs are correlated with emotional evaluation indices. Interestingly, while the satisfaction levels of autonomy and competence correlate with both positive and negative emotions, the relatedness level correlates only with positive emotions. Closely related persons evoke more emotions than strangers. If relatedness is not satisfied, unpleasant emotions are not necessarily evoked, but people sense discomfort when an expected reaction is not delivered by the related person. Thus, compared with the other two needs, relatedness exerts a more complicated effect on emotions.

The need for relatedness becomes apparent from the early stage of infant development. Still-face paradigm experiments have shown that infants are socially sensitive to others (Adamson and Frick, 2003; Striano, 2004). In these experiments, the caregiver

suddenly ceases normal interaction with the infant and shows a still face. Throughout this phase, the caregiver reduces the number and extent of positive activities, such as smiles or attention. Infants react to this behavior with restore reactions such as clapping or reaching to the caregiver to draw their attention. The reactions shown by infants depend on their development stages (Adamson and Frick, 2003). These experiments show that infants are motivated to establish relatedness with others and that they require attachment to others.

Several studies in cognitive developmental robotics (Asada et al., 2009) have sought to understand initial communication by computational models (Ogino et al., 2007; Watanabe et al., 2007). However, these studies focused on acquiring communicative actions, rather than the factors that motivate communication. The mechanism that encourages an agent to behave according to internal discipline rather than external reward has been identified as intrinsic motivation (Barto et al., 2004; Oudeyer et al., 2007). However, intrinsic motivation studies continue to adopt self-learning tasks such as skill acquisition. The question remains: how do intrinsic motivation mechanisms promote communicative interactions?

This paper proposes a motivation model of early communication between an infant and his/her parent, in which the need for relatedness triggers emotional change and behavior learning. The proposed model aims for dynamic interaction between two agents who estimate each other's internal state. Throughout the interaction, an interpersonal relationship is established in which approaching and sharing another's emotion encourages interest and relatedness to him/her, alters emotional states, and promotes mutual behavior. While relatedness directly and indirectly affects the emotional state of an agent, it also changes the reward for action selection. We consider that the network of dopamine

neurons play an important role in activating communication. Dopamine neurons are known to code the prediction error for reward in reinforcement learning (Schultz, 1998). In robotics, (Kaplan and Oudeyer, 2007) hypothesized that dopamine neurons encode signals for encouraging behavior that decreases the prediction error. Recent studies reveal that dopamine neurons are activated not only by explicit reward but also by novel signals that are not directly related to these rewards (Dayan and Balleine, 2002; Kakade and Dayan, 2002). This indicates that dopamine neurons play an important role in intrinsic motivation. Dopamine neurons are also associated with emotional reactions in the amygdala (Phillips et al., 2010). From these neuroscience findings, it is reasonable to consider a model in which a variable corresponding to dopamine neurons mediates emotional change and behavior. In parent–infant interactions, the activation of dopamine neurons will arouse the infant's interest, and the parent will act to maintain this interest.

2. MATERIALS AND METHODS

2.1. MOTIVATION MODEL OF PARENT-INFANT INTERACTION

In the communication situation of this study, an infant attentively interacts with his/her caregiver and displays emotional facial expressions such as laughing and crying. The interaction situation and variables used in the proposed model are shown in **Figure 1A**. The infant and the parent update their internal state, e , based on the observed information, x , and output their facial expressions, f , and actions, a . The facial expressions and actions are assumed to be produced and observed independently. The facial expressions are based on the agent's internal state, e , which partly depends on the facial expressions of the other agent, e_{other} . We suppose that both agents (parent and child) possess the same emotional system, comprising *emotional elements*, *emotion*, and *action selection* modules (**Figure 1B**). The *emotional elements* module contains two main elements for intrinsic needs, *Novelty* and *Relatedness*, and other three sub-elements, *Stress*, *Emotion Mirror* and *Expectation*. The value of each element is determined by the other's facial expressions and actions. The emotion elements are used to compute the current emotional state of the agent in the *emotion module*. Finally, in the *action module*, the reward value is evaluated from the emotional elements (pleasure and arousal), and gesture and facial expressions are selected. The following subsections describe the mechanisms of the internal state.

2.1.1. Emotion

Russell (1980) proposed that all emotional states lie within a two-dimensional space comprising an arousal–sleep axis and a pleasure–unpleasantness axis. Following Russell's model, we define the emotional state e as a vector of arousal and pleasure elements.

$$e(t) = \begin{pmatrix} e^{\text{Arousal}}(t) \\ e^{\text{Pleasure}}(t) \end{pmatrix} = \begin{pmatrix} e^A(t) \\ e^P(t) \end{pmatrix} \quad (1)$$

The emotional state is updated by the reward, R_e , as follows;

$$e(t+1) = e(t) + \eta(R_e(t) - e(t)). \quad (2)$$

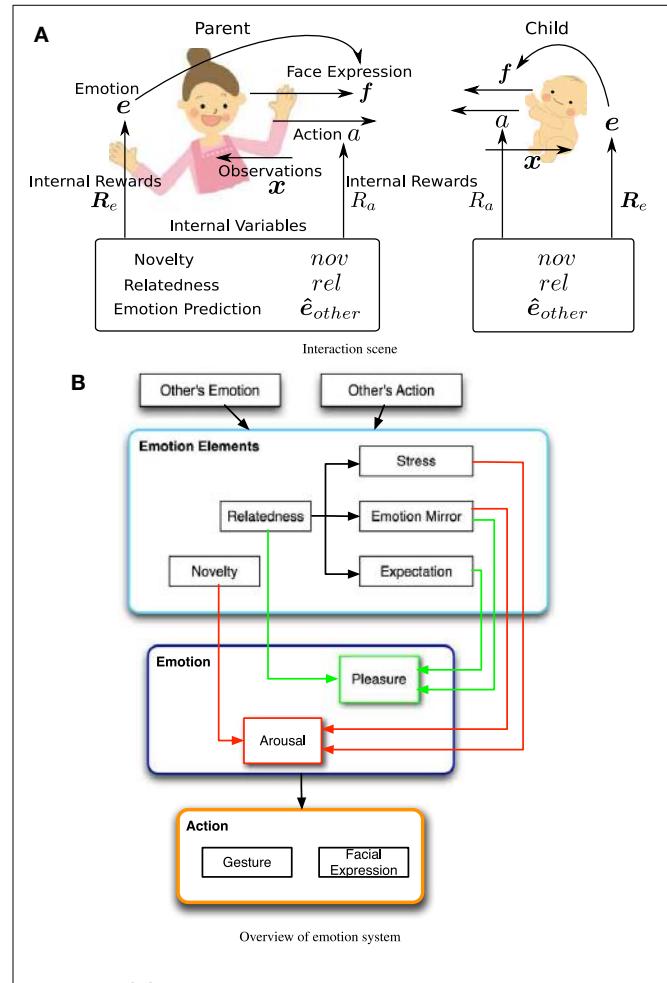


FIGURE 1 | (A) Variables used in the proposed model in an interaction situation. **(B)** Overview of the emotion system.

The elements of the reward corresponding to arousal and pleasure, $R_e^A \cdot R_e^P$, are composed of various psychological factors—novelty, relatedness, emotional contagion, and expectancy—denoted nov, rel, e_{const} , str, and E_{grad} , respectively.

$$R_e^A(t) = \alpha^A \text{nov}(t) + \beta^A \text{str}(t) + \gamma e_{\text{cont}}^A(t) \quad (3)$$

$$R_e^P(t) = \alpha^P \text{rel}(t) + \beta^P E_{\text{grad}}(t) + \gamma e_{\text{cont}}^P(t). \quad (4)$$

The novelty, nov, indicates the degree of interest in novel surrounding objects, and it is defined as

$$\text{nov}(t) = 1/(1 + \exp(-m(I(t) - \theta))), \quad (5)$$

where $I(t)$ is information gain, and m and θ are constants. The information gain, $I(t)$, is based on a state transition model constructed by the agent's observations,

$$I(t) = -\log p(s(t+1) | s(t)). \quad (6)$$

Stern (1985) proposed that the emotional attunement of a parent is important in establishing a parent–child relationship. Such

emotional attunement is thought to be necessary for the sharing mind states. Thus, we assume that the relatedness variable, rel , depends on the synchronization of emotional states:

$$\text{rel}(t) = (1 - \mu)\text{rel}(t - 1) + v\text{sim}(t), \quad (7)$$

where μ and v are constants. sim is the emotional similarity, i.e., the extent to which other's emotional states are shared between the agents. sim is the inner product of the self and other's emotion vectors:

$$\text{sim}(t) = e(t)e_{\text{other}}(t). \quad (8)$$

As suggested by fMRI experiments (Singer et al., 2004), humans possess an emotional mirror system. A person's emotional state is slightly altered by the perceived emotional state of another. In this paper, the variable for emotional contagion variable, e_{cont} , is the product of relatedness and the emotional state of the other:

$$e_{\text{cont}}(t) = \text{rel}(t)e_{\text{other}}(t). \quad (9)$$

Note that the emotional contagion increases with increasing degree of relatedness.

When a parent is unwilling to relate to his/her infant, the heart rate of the infant increases, and the infant's gaze is averted from the parent, apparently because the infant is temporarily aroused by the stress of communication failure (Field, 1981). In our model, the stress variable increases when emotional sharing with the related person fails; that is

$$\text{str}(t) = \text{rel}(t) \exp(-\sigma\text{sim}(t)). \quad (10)$$

where σ is a positive constant.

While emotional contagion and stress cause temporary effects, the impact of emotional expectancy is long lasting. For example, pleasure is enhanced when one's action appears to please another. Thus, we define emotional expectancy as temporal gradient of expected pleasure, defined by multiplying the action selection probability by the pleasure of the other at the present and preceding moments, and taking their difference.

$$E_{\text{grad}}(t) = \text{rel}(t) (p_a(t)e_{\text{other}}^P(t) - p_a(t-t_b)e_{\text{other}}^P(t-t_b)) \quad (11)$$

The emotional expectancy is large when the action selection probability and the pleasure emotion increase together. Emotional expectancy is also affected by relatedness.

2.1.2. Motivation mechanism for action

Dopamine neurons in the midbrain are considered to encode values; they are activated and suppressed in desirable and undesirable situations, respectively. However, some dopamine neurons have recently been reported as activated even in undesirable situations. Bromberg-Martin et al. (2010) proposed that dopamine neurons encode either motivational value or motivational salience. Thus, we model two classes of dopamine neurons, as follows.

Dopamine neurons belonging to the first class, encoding a motivational value, are projected from the basal ganglia and contribute to the exploratory and evaluative learning of whether the current situation is desirable/undesirable. In infant-parent communication, actions that attract the infant's interest and establish relatedness will score high motivational value. Thus, we suppose that the first class of dopamine neurons encodes the other's arousal emotion and relatedness,

$$R_a^{\text{Value}}(t) = \hat{e}_{\text{other}}^A(t) + \omega \text{rel}(t), \quad (12)$$

where ω is a positive constant.

Dopamine neurons belonging to the second class, motivational salience, are projected from the amygdala. The neurons contribute to the learning of motivationally important events that may not be related to reward and are thought to aid attention and working memory. We suppose that the second class of dopamine neurons encodes the arousal emotion

$$R_a^{\text{Salience}}(t) = e^A(t). \quad (13)$$

Both rewards are summed to give the total reward

$$R_a(t) = \rho R_a^{\text{Value}}(t) + (1 - \rho) R_a^{\text{Salience}}(t), \quad (14)$$

where ρ is a weighting constant ($0 \leq \rho \leq 1$). As ρ increases, an agent acts upon predictions of the other's emotional state. If ρ is small, an agent acts more upon its own emotional response.

Reinforcement learning is used to update the action policy. Although various sensor information is important in actual communication, here we consider actions alone. When an action a yields a reward R_a , the corresponding action value function \hat{R}_a is updated as

$$\hat{R}_a(t+1) \leftarrow \hat{R}_a(t) + \eta_{R_a} (R_a(t) - \hat{R}_a(t)), \quad (15)$$

where η_{R_a} is a learning coefficient.

The parent agent adopts an ε -greedy policy. That is, with probability ε , the parent agent selects the highest-valued action, \hat{R}_a , among its own action repertoire, and otherwise chooses random actions,

$$\pi_{\text{parent}}(t) = \begin{cases} \text{random action} & (\zeta < \varepsilon) \\ \arg \max_a \hat{R}_a(t) & (\text{otherwise}) \end{cases}, \quad (16)$$

where ζ is randomly drawn from a uniform distribution in the interval $[0, 1]$.

The actions performed by the infant agent depend on the action value \hat{R}_a . The movements of infants appear random and occasional, whereas those of their parents are voluntary. Thus, we model the selection and performance of infant actions by a Boltzmann equation

$$\pi_{\text{child}}(t) = \frac{\exp(\hat{R}_a(t)/\tau)}{c + \exp(\hat{R}_a(t)/\tau)}, \quad (17)$$

where c is the initial probability that an action is taken. The temperature parameter τ determines the randomness of action selection.

2.2. INTERACTION EXPERIMENT WITH A VIRTUAL ROBOT

To validate its applicability in real communication, the proposed model was implemented in a virtual agent. The virtual agent communicates with a human experimenter who mimics parent-like facial expressions and behavior.

The experimental setup is shown in **Figure 2**. Displayed on a laptop computer, the virtual agent (**Figure 2A**) observes facial expressions and behavior of the experimenter by a USB camera attached to the top of the laptop display.

The virtual agent displays four types of facial expressions depending on its emotional state. It also exhibits an appealing behavior by arm movement. The facial expressions and appealing behavior are shown in **Figure 3**.

The experiment was undertaken in two phases. In the first (learning) phase, the virtual agent learns the relationship between the experimenter's facial expressions and its corresponding emotional states and constructs a layered network for behavior recognition. In the second (interaction) phase, the virtual agent communicates with the experimenter.

In the learning phase, information for emotional estimation and behavior detection is processed from camera images. During

emotional estimation, the estimated emotional state of the experimenter, e_{other} , is output from the camera image, x . The facial area in the captured image is extracted by the facial recognition algorithm in OpenCV (Bradski, 2000), converted to a gray scale image of size 128×128 pixels, and binarized by a specified threshold. In the learning phase, a certain number of facial images, I_i , is recorded and each is stored with its corresponding emotional state, e_i . The correspondence between the emotions of the virtual and human agents is learned by imitation (Watanabe et al., 2007); that is, the human agent imitates the facial expressions of the virtual agent when presented with a stimulus such as a blue object or keyboard pressing (these responses of the virtual agents are pre-programmed). In the interaction phase, the input facial image I_x is compared with the stored images and the best-matched facial image is selected as

$$I_{\min} = \arg \min_{I_i} |I_x - I_i|^2 \quad (18)$$

Let ψ be the mapping function. The momentary estimated emotional state of the experimenter is calculated as

$$e_{\text{now}} = \psi(I_{\min}). \quad (19)$$

The estimated emotional state is the temporal average of the momentary estimated emotional states,

$$e_{\text{other}} = (1 - \delta)e_{\text{other}} + \delta e_{\text{now}} \quad (20)$$

where δ is an update constant.

Figure 4 shows the learned map of the facial images and their corresponding emotional states. The vertical and horizontal axes indicate the arousal and pleasure levels, respectively. For this experiment, the emotional state of the experimenter is estimated from 25 images.

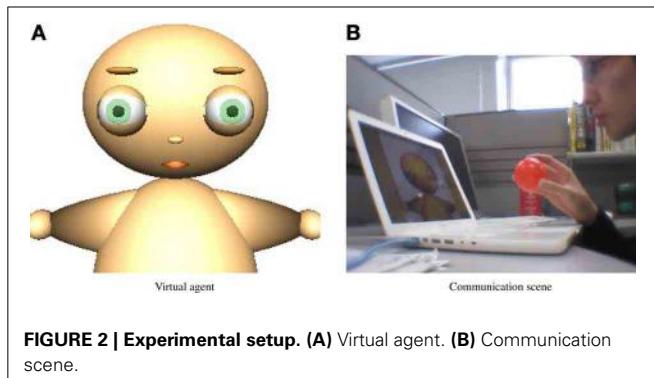


FIGURE 2 | Experimental setup. (A) Virtual agent. (B) Communication scene.

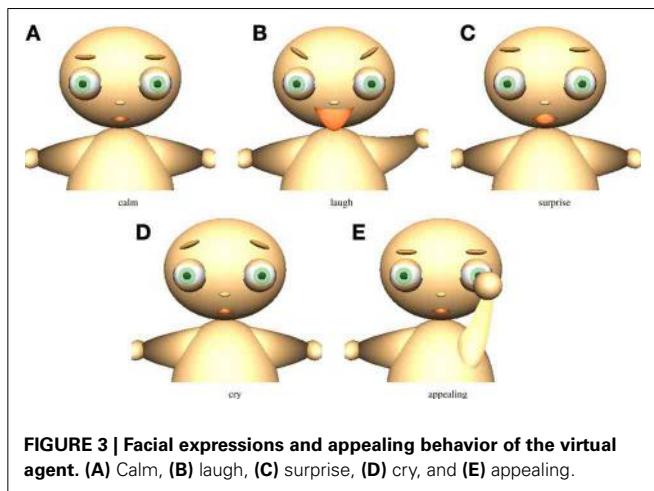


FIGURE 3 | Facial expressions and appealing behavior of the virtual agent. (A) Calm, (B) laugh, (C) surprise, (D) cry, and (E) appealing.

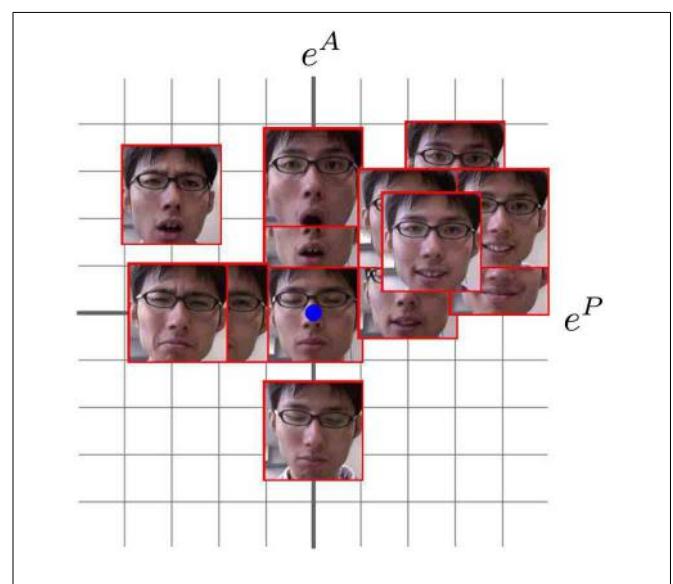


FIGURE 4 | Learned map of facial images and emotional states.

Behavior recognition is achieved by a layered neural network of slow feature analysis (SFA). The SFA learning algorithm extracts the slowly changing components from the input signals and estimates the inherent information based on their statistical properties (Wiskott and Sejnowski, 2002). According to some studies, SFA exhibits stronger gesture recognition performance than existing methods such as the hidden Markov model and random forest (Koch et al., 2010).

The input to the SFA layered network is an image of the experimenter waving a red object in his hand in various directions. The learning data are 2000 steps of images. The input image (320×240 pixels) is segmented into the small areas of size 16×12 pixels. Each small area is labeled as “1” if the number of red pixels (specified by RGB content $R \geq 160, G \leq 50, B \leq 80$) exceeds half; otherwise, it is labeled “0”. The resultant 128-dimensional vector is used to construct a state transition model. The range of the j -th unit in the SFA output layer, y_j , is divided into S_j bins. The output signal is described by the discrete states $s(y_j)$ ($s \in \{1, 2, \dots, S_j\}$), from which the state transition probability in the j -th output signal, y_j , is calculated as

$$p_{ss'}^j = \Pr\{s(y_j(t+1)) = s' | s(y_j(t)) = s\}. \quad (21)$$

This state transition model is iteratively updated when a new state is observed.

The information gain of y_j , $I_j(t)$, is calculated by the state transition model as

$$I_j(t) = -\frac{1}{t_a+1} \sum_{t=t-t_a}^t \log p(s(y_j(t+1)) | s(y_j(t))). \quad (22)$$

From 22, the novelty of the j -th output signal is evaluated as

$$\text{nov}_j(t) = \frac{1}{1+\exp(-m(I_j(t)-\theta))}. \quad (23)$$

Finally, the novelty of the whole output signal, $\text{nov}(t)$, is calculated as the average of the novelty of each output signal

$$\text{nov}(t) = \frac{1}{n} \sum_j^n \text{nov}_j(t). \quad (24)$$

During the interaction phase, the experimenter communicates with the virtual infant agent with a red object in his/her hand. The communication mimics the still-face paradigm experiment

in developmental psychology, passing through the three phases of interaction, still face, and reunion. During the interaction phase, the experimenter looks at the camera and expresses surprise, simulating a parent seeking the attention of his/her infant. Then, when the agent similarly expresses surprise, the experimenter ensures that the arousal emotion is shared and begins laughing to the virtual agent. Throughout the interaction, the experimenter moves the red object, starting with the action patterns shown in Figure 5, and later by free motion. The persistent changes in the action pattern maintain the arousal level and the attention of the virtual infant.

During the second phase (still face), the experimenter ceases object movement and shows a blank facial expression. The possible unfamiliarity between experimenter and agent is non-problematic, because in actual still-face paradigm experiments, the still-face effect is elicited in infants meeting a person for the first time (Adamson and Frick, 2003).

During the last phase (reunion), the experimenter reverts to the interaction phase; that is, moving the red object and presenting emotional facial expressions.

The virtual agent shows simulations laughing ($e^P > 0.4$), crying ($e^P < 0$), surprise ($0 < e^P < 0.4$, $e^A > 0.4$), and normal (otherwise). The relationship between the emotional states and

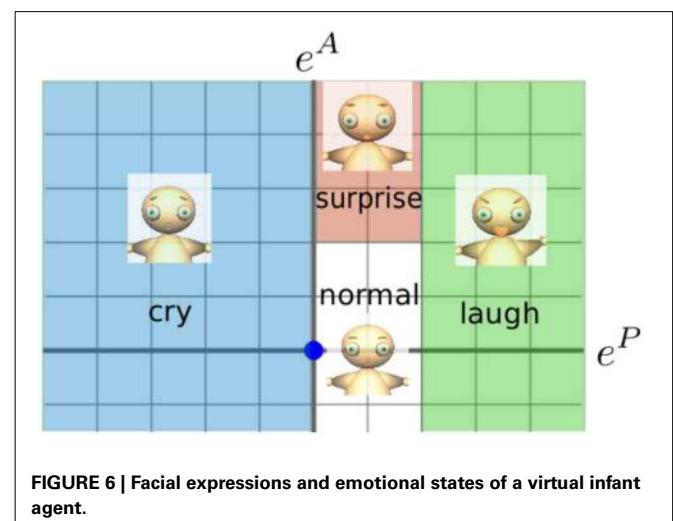


FIGURE 6 | Facial expressions and emotional states of a virtual infant agent.

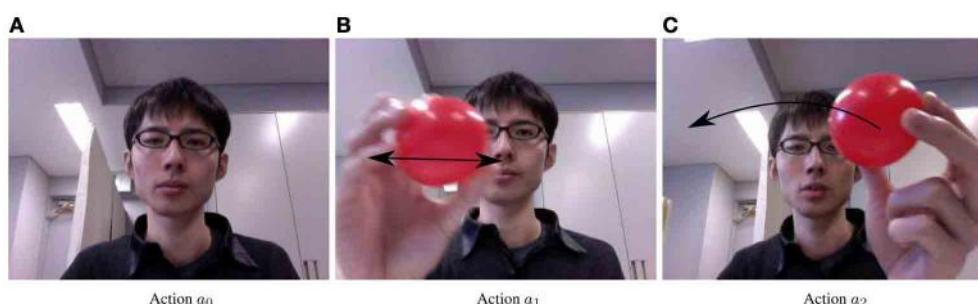


FIGURE 5 | Actions made by experimenter. (A) Action a_0 . (B) Action a_1 . (C) Action a_2 .

the facial expressions is shown in **Figure 6**. The agent appeals (**Figure 3E**) to the experimenter based on the probability of action taken (Equation 17).

The simulated still-face paradigm experiment was conducted over the time frame of the equivalent developmental psychology experiment (Adamson and Frick, 2003); 2 min for the first interaction, 2 min for the still face, and 2 min for the reunion.

While the emotional expressions and actions of the human experimenter are continuous, they are undertaken by the virtual agent in a numerical time step (430 ms). To maintain natural interaction, the agent retains the triggered facial expression and action for 4 s. The experimenter's emotional state is estimated every 5 steps. The emotional state of the agent and all other variables are updated at each step. The novelty is evaluated after each 30-step sequence of human actions (i.e., the t_a is set to 30 in Equation 22). The constants and parameter settings of the infant agent are described in the next section.

2.3. SIMULATION EXPERIMENT OF PARENT-INFANT INTERACTION

We suppose that parent–infant interaction is enabled by a motivation mechanism that is common to both individuals. The communication dynamics are governed by mutual attraction between the infant's and parent's motivation. In this section, the proposed model is implemented in two agents to determine whether the parent–infant interaction emerges through interplay between emotion and action in a simulation environment. We also examine how the relatedness of the infant agent changes in response to varying patterns of action and emotional expressions presented by the parent agent.

Both agents are assigned three actions, a_0 , a_1 and a_2 , as shown in **Figure 5**. The parent agent selects its action from the repertoire when the action value is updated. If the probability of action (Equation 17) exceeds a given threshold, the infant agent adopts the action taken by the parent in the previous step.

Action recognition is based on the image sequence recorded in the interaction experiment between the human and the virtual agent. When its partner performs an action, the observing agent accepts an image sequence (30 images 9 of the active agent as input. When no action of the agent is observed, the novelty of the observer decreased by a factor of λ ,

$$\text{nov}(t) = \lambda \text{nov}(t - 1) \quad (\text{if no action is observed}). \quad (25)$$

In emotional expression and recognition, we assume for simplicity that one agent can observe the emotion of another agent, e_{other} , from his/her facial expression, f_{other} without mistakes.

We also assume that two steps of simulation time correspond to 1 s. An action is selected, and the action value, together with the emotional estimate of another agent, is updated every 5 steps. All other variables, including the emotional state, are updated at each step.

We allocated the following five conditions of facial expression and action patterns of a parent agent.

1. normal
2. still face

3. fixed action
4. random emotion/action
5. no relatedness

Under condition (1) *normal*, the parent agent behaves according to the proposed emotional system.

Under condition (2) *still face*, the parent agent adopts the still-face behavior in human-agent interaction experiments. The simulated experiment is undertaken in three phases; interaction phase (0–999 steps), still face phase (1000–1199 steps), and reunion phase (1200–1399 steps). Each phase corresponds to 2–3 min in real time. While both agents follow the proposed model during the interaction phase, the parent ceases facial expression and activity in the still-face phase. During this phase, the emotional state of the parent agent is set to $e^A = 0$ and $e^P = 0$. In the reunion phase, the parent agent recovers its emotional expression and resumes action.

Under condition (3), *fixed action*, the parent agent selects the same action, a_1 , while its emotional expressions are governed by the proposed model. Unlike the normal condition, in which action selection by the parent depends on the action value, \hat{R}_a , the fixed action arouses marginal emotion in the infant. The resulting lack of novelty perceived by the infant reduces the relatedness.

Under condition (4), *random emotion/action*, the parent expresses random emotion expressions and performs actions randomly. The arousal and pleasure values are randomly selected from −1 to 1. Among the three-action repertoire, each action is selected with equal probability. While emotions are continuously shared between the parent and infant agents under normal conditions, emotional sharing is interrupted under this condition.

Under condition (5), *no relation*, the relatedness of the parent agent is not updated (and remains fixed at 0). This condition enables the observation of how relatedness between the agents affects their emotional sharing.

The parameters and coefficients used in this experiment are listed in **Tables 1–4**.

3. RESULTS

3.1. EXPERIMENTAL RESULTS IN INTERACTION EXPERIMENT WITH VIRTUAL ROBOT

Throughout the 6-min interaction period, the virtual agent completed 828 calculation steps. **Figure 7** shows the temporal profiles of relatedness during the interaction. Throughout the first interaction phase, the relatedness increases to its maximum value 1.0 in 118 s. The relatedness declines throughout the still-face phase (from 120 to 240 s) is minimized (0.33) at 247 s and

Table 1 | Parameters of emotional elements.

Parameter	Explanation	Parent/ Infant
m	Coefficient of information gain for novelty	100
θ	Threshold of information gain for novelty	0.9
μ	Decay constant for relatedness	0.006
v	Coefficient of vector similarity for relatedness	0.025
σ	Coefficient of similarity for stress	5

Table 2 | Parameters of emotional change.

Parameter	Explanation	Parent	Infant
α^A	Coefficient of novelty for arousal reward	0.8	0.5
β^A	Coefficient of stress for arousal reward	0	2
α^P	Coefficient of relatedness for pleasure reward	0.8	0.45
β^P	Coefficient of expectancy for pleasure reward	0	40
γ	Coefficient of emotional contagion for pleasure reward	0.6	1
η_{Re}	Coefficient for update of emotional state		0.03

Table 3 | Parameters of action motivation.

Parameter	Explanation	Parent	Infant
ω	Coefficient of relatedness for motivational value	0.3	–
ρ	Weight of motivational value for action reward	1	0
η_{Ra}	Coefficient of action value update		0.6
ε	Probability that parent selects random action	0.1	–
c	Initial constant of action selection of infant	–	4
τ	Temperature constant of action occurrence probability of infant	–	0.3

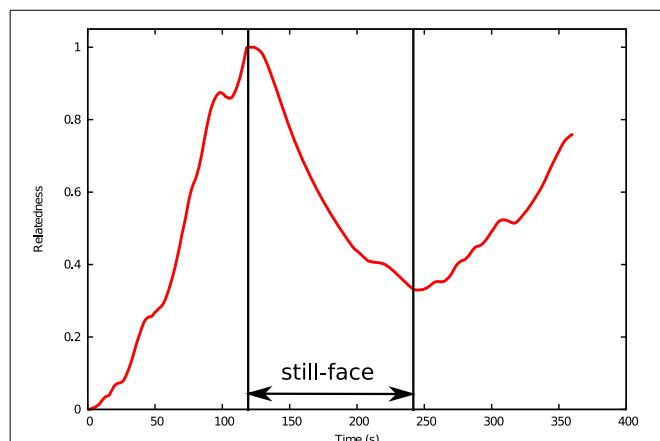
Table 4 | Other system parameters.

Parameter	Explanation	Parent/ Infant
n	Number of input signals for novelty detection	20
S_j	Number of bins in input signals for novelty detection	20
η_e	Coefficient for emotion estimation	0.05

recovers throughout the reunion phase (after 240 s) when normal interaction is resumed.

Figure 8A shows the emotional state of the experimenter estimated by the infant agent. While the experimenter shows a positive emotional state in the interaction and reunion phases, its arousal and pleasure value fall to 0 during the still-face phase.

Figure 8B shows how the emotional state of the infant virtual agent changes over time. During the first interaction phase, positive emotion continues, and the pleasure level increases with increasing relatedness. Note that the arousal level suddenly escalates in the still-face phase, while the pleasure level decreases. During the reunion phase, the arousal settles around 0.5, and the pleasure recovers. **Figure 9** shows the probability of action taken by the agent. This probability increases with increasing pleasure level throughout the interaction phase, but suddenly leaps in the still-face phase. This trend mirrors the appealing behavior of infants real-time still-face experiments.

**FIGURE 7 | Relatedness of virtual agent as a function of time in the simulated still-face experiment.**

3.2. EXPERIMENTAL RESULTS OF SIMULATED PARENT-INFANT INTERACTION

Figure 10 shows the temporal dynamics of relatedness in the infant agent while interacting with the parent agent under the five conditions. While the relatedness increases to its maximum as the interaction proceeds under *normal* and *still-face* conditions, it remains low under the remaining three conditions. During the 860 steps of the interaction phase under the *still-face* condition (corresponding to the *normal* condition), the relatedness increases to 1. However, while the relatedness remains at 1 under normal conditions, it declines throughout the still-face phase, because the parent shows no emotional expression, and the degree of emotional sharing, sim, reduces to 0. In the reunion phase, after 1202 steps, the relatedness recovers as observed in the human–robot interaction experiment.

Figure 11 shows the emotional states of infant and parent agents. Throughout the interaction phase, the actions of the parent engage the infant agent, raising its arousal level. The increased relatedness enhances the pleasure level in both agents. During the still-face phase, both the arousal and pleasure levels of the parent agent decrease to 0 (**Figure 11A**). On the other hand, the resulting stress to the infant (described by Equation 10) increases its arousal level (**Figure 11B**). Increased arousal is accompanied by a decline in the pleasure level shortly after entering the still-face phase. This negative emotion is induced by the negative value of expectancy value (Equation 11). During the reunion phase, the arousal level of the infant decreases to pre-stress levels, and the pleasure level is recovered as relatedness is restored.

Under the *fixed action* condition, the relatedness increases to 0.1 and gradually declines to a low level. By contrast, relatedness remains low under *random action/emotion* conditions. As defined in Equation (7), relatedness is determined by the similarity of emotional states between the two agents. Throughout the interaction phase, the arousal level of both agents necessarily increases, as specified in the still-face condition. However, when the parent performs fixed actions, it stimulates no novelty in the infant. Although random actions do stimulate novelty in

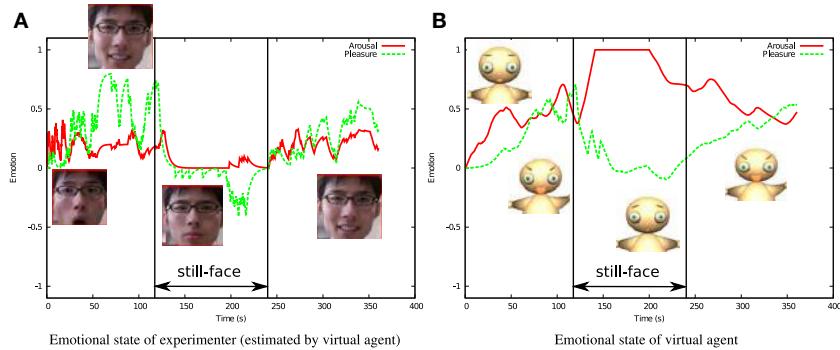


FIGURE 8 | Emotional state of experimenter (estimated by virtual agent) (A) and virtual agent (B).

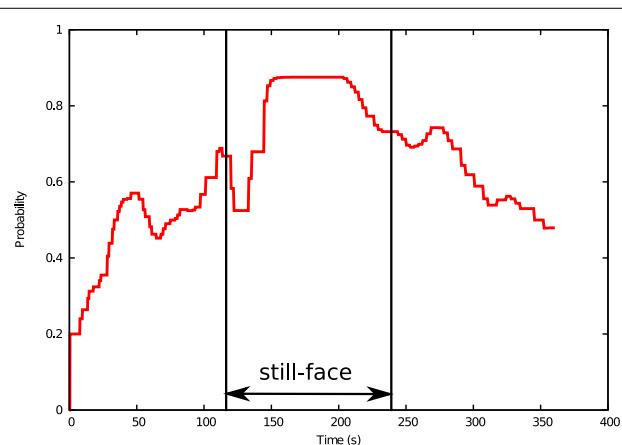


FIGURE 9 | Probability of agent action as a function of time in the simulated still-face experiment.

the infant, the randomness of the parent's emotional expressions interrupts emotional sharing, thereby reducing the relatedness under *random action/emotion* conditions.

Under the *no relatedness* condition, the relatedness of the infant agent increases up to around 0.2 during the first 400 steps and remains at 0.2 thereafter. During the interaction phase, the shared arousal emotion enhances the relatedness. Subsequently, the shared pleasure emotion further increases the similarity, sim, and thus the relatedness. However, since relatedness dominates the pleasure level, pleasure cannot increase if the parent lacks relatedness. Thus, high relatedness in the infant agent can be achieved only by the sharing of arousal emotion.

4. DISCUSSION AND CONCLUSION

In our model, parent infant interactions are primarily mediated through novelty and relatedness. Novelty motivates interaction with the environment. Since the novelty value is evaluated from a pre-learned state transition model, it is increased by the perception of dynamic movement and reduced in still environments. Based on this property, the parent predicts which action will elicit higher novelty in an infant, such as moving an object.

As the infant detects novelty in his/her parent's behavior, its arousal level and response frequency are enhanced. In turn, the infant responses evoke novelty, and hence arousal, in the parent. Increased arousal in both agents increases emotional sharing, and hence the relatedness, between the agents. This enhanced relatedness encourages pleasurable emotions and further emotional sharing. The simulation experiment demonstrated this positive feedback effect of mutually exchanged rewards.

In the proposed model, novelty and the state transition probabilities of other agent's actions are evaluated by SFA networks. Such networks are effective for extracting similar action structures from image sequences, because they can integrate temporarily similar information. This property of SFA networks renders them suitable for gesture recognition, where repeat observations of the same action are perturbed by human motion and lighting conditions. In fact, unvarying repeated action decreases the novelty, because the same state transition is observed.

The relatedness modeled in this paper does not consider long-term relationships. We reiterate that the still-face paradigm is applicable not only to parent–infant interactions but also to stranger–infant interactions (Adamson and Frick, 2003). Furthermore, the still-face response is absent during interactions with impersonal objects. This finding indicates that infants can relatively quickly identify whether an object/person is amenable to social interaction and can relate to that object or person. Humans do not empathize with objects and other humans that fail to comply with expectation, unless relatedness is also present. Relatedness is regarded as a precursor to all social emotions, including social expectation, social contagion, and social stress. For this reason, the modeled terms of social contagion, stress, and expectation of emotional reward include multiples of relatedness.

An interesting result of the proposed model is that surprise appears first in the interaction, followed by pleasure. This is attributable to the evocation of arousal by novelty detection, which occurs regardless of relatedness, while the pleasure emotion arises only through relatedness. Thus, during the initial interaction, when relatedness is low, arousal is elicited first. Next, as arousal is shared, the relatedness is increased, followed by pleasure, which elicits the smiling response. In this way, emotional contagion encourages further emotional sharing.

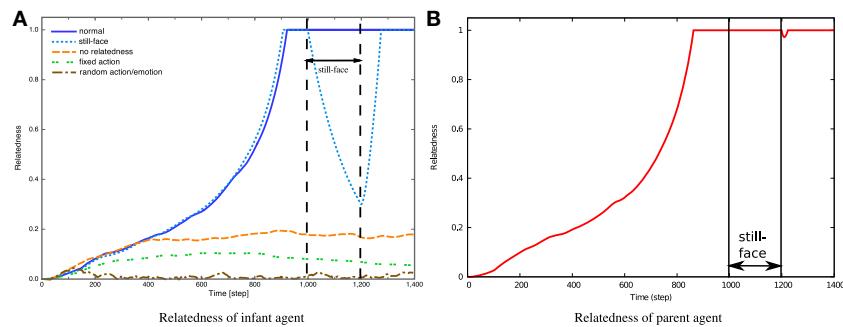


FIGURE 10 | Relatedness during simulated parent–infant interactions. (A) Relatedness of infant agent. **(B)** Relatedness of parent agent.

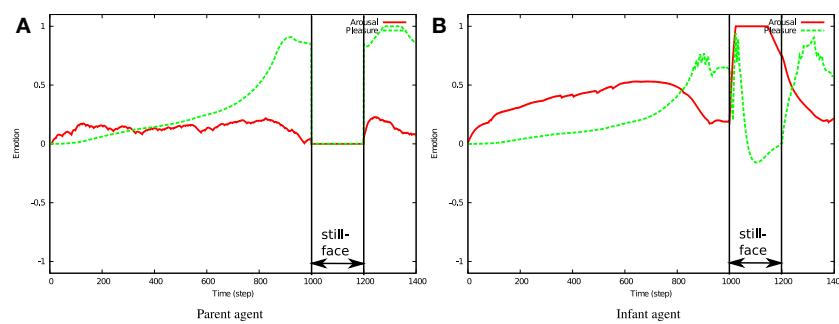


FIGURE 11 | Emotional states during simulated still-face interactions. Red and green lines indicate the arousal and pleasure levels, respectively. **(A)** Parent agent. **(B)** Infant agent.

Although the parent frequently changes action during interaction phase, the frequency of change decreases as relatedness increases. High relatedness maintains the motivation at a high level and prevents the decline of the action value. Indeed, in the simulation experiment, the parent altered its actions 41 times through the interaction phase, increasing to 96 times when relatedness was set to 0. In actual human communications, this trend might signify a shift from a unidirectional form, in which a parent attracts the attention of an infant, to a bidirectional form, in which both parent and child pursue pleasurable emotions.

The simulation experiment investigated how the interaction between the parent–infant interaction changes when the relatedness of the parent agent is not updated. Under this condition, the emotional state of the parent is static, and the action patterns depend on emotional sharing with the infant. During the first phase, the arousal level increases in both parent and infant agents, increasing the relatedness and pleasure levels of the infant, while those of the parent remain fixed. In this case, because the emotional state vectors of both agents diverge, the relatedness and pleasure levels of the infant remain low. Thus, if one agent seeks relatedness and its accompanying pleasure, it must find another agent with the same goal at the same time. Baumeister and Leary (1995) proposed that human beings are fundamentally and pervasively motivated by a need to belong; that is, to form enduring interpersonal attachments. According to these authors, this need is satisfied when pleasant interactions occur within a

temporally stable and enduring framework of affective concern for each other's welfare. In our simulation study, a similar reciprocal relationship between two agents was required to maintain interpersonal attachments.

In the simulation experiment, the relatedness was initialized to 0 both in both agents. In an actual interaction, the parent who establishes communication with his/her infant possesses high relatedness at the beginning of the interaction. However, if the initial pleasure value of the parent agent is set to 1, the relatedness decreases, because the pleasure level does not match that of the infant. This occurs because relatedness in the proposed model depends only on emotional similarity. This problem might be solved by including a top–down mechanism, such as a bias term, when calculating the relatedness in the parent agent. Such a term would account for the parent's desire to interact with the infant.

In this paper, the emotional state of an agent is defined in a two-dimensional plane whose axes are arousal and pleasure. This low-dimensional model of emotions has been previously adopted in robotics studies (Breazeal and Scassellati, 1999; Itoha et al., 2005; Watanabe et al., 2007). In psychology, low-dimensional models are based on descriptive taxonomies and have proven reasonably successful for describing measures of self-reported emotion and relative confusion of various facial expressions. However, the sections of brain corresponds to each dimension are not clear. Arguably, such models cannot explain selective emotional

impairments (Calder et al., 2001). The difficulties in modeling emotions necessitate a direct quantitative comparison of the model with psychological experiments. Facial expressions and physiological data such as Galvanic skin response are superficial expressions of internal emotional states. In this paper, the still-face effect is qualitatively compared with the psychological still-face paradigm experiment. Although emotions appear to be dispersed within the human brain, unlike the physical sense of touch, which is located in the somatosensory area, separated areas are probably connected within the state space of emotion. In future experiments, we plan to incorporate brain mechanisms, including the relationships among brain regions related to emotion, and to compare the theoretical model with brain activities during interactions (Dumas et al., 2010).

Gaze is one of the most important challenges in extending the proposed model. Arousal is closely related to attention. In the proposed model, an agent informs its interest to another by arousal-induced actions but does not inform the item of interest. Furthermore, the parent's action value varies over time, but it is independent of sensor information. Supplied with gaze information, a parent could locate and identify the item commanding the infant's attention, which would enrich communication. For example, parental behavior such as intentionally shifting the timing of an action or showing exaggerated facial expressions after

attracting the infant's attention would further enhance pleasure in the infant.

The proposed model does not explain the decrease of the infant's attention toward the parent in the still-face phase. Such behavior is thought to decrease the stress experienced by the infant (Field, 1981). If true, our model must introduce attention mechanisms for controlling emotion. Furthermore, including gaze information, we could extend our simulated interactions from dyadic interactions to triadic relationships among parent, infant, and object. Especially, joint attention, in which the infant attends to the object occupying the parent's attention or promotes the parent to attend to his/her object of interest, is an important topic in the communication of shared emotion. The learning of joint attention has already been modeled in developmental cognitive robotics (Nagai et al., 2003; Triesch et al., 2006). In future extensions of our model, we aspire to understand how higher cognitive functions such as joint attention relate to motivational behavior such as novelty and relatedness.

ACKNOWLEDGMENTS

This study was supported by a Grant-in-Aid for Specially Promoted Research, JSPS KAKENHI Grant Number 24000012. The authors would like to thank Matthias Rolf and Enago (www.enago.jp) for the English language review.

REFERENCES

- Adamson, L. B., and Frick, J. E. (2003). The still face: a history of a shared experimental paradigm. *Infancy* 4, 451–473. doi: 10.1207/S15327078IN0404_01
- Asada, M., Hosoda, K., Kuniyoshi, Y., Ishiguro, H., Inui, T., Yoshikawa, Y., et al. (2009). Cognitive developmental robotics: a survey. *IEEE Trans. Auton. Ment. Dev.* 1, 12–34. doi: 10.1109/TAMD.2009.2021702
- Barto, A. G., Singh, S., and Chentanez, N. (2004). “Intrinsically motivated learning of hierarchical collections of skills,” in *Proceedings of the 3rd International Conference on Developmental Learning* (San Diego, CA), 112–119.
- Baumeister, R. F., and Leary, M. R. (1995). The need to belong: desire for interpersonal attachments as a fundamental human motivation. *Psychol. Bull.* 117, 497–529. doi: 10.1037/0033-295X.117.3.497
- Berlyne, D. (1960). *Conflict, Arousal and Curiosity*. New York, NY: McGraw-Hill. doi: 10.1037/11164-000
- Bradski, G. (2000). The OpenCV library. *j-DDJ*, 25, 120, 122–125.
- Breazeal, C., and Scassellati, B. (1999). “How to build robots that make friends and influence people,” in *Proceedings of the IEEE International Conference on Intelligent Robots and Systems* (Kyongju), 858–863. doi: 10.1109/IROS.1999.812787
- Bromberg-Martin, E. S., Matsumoto, M., and Hikosaka, O. (2010). Dopamine in motivational control: rewarding, aversive, and alerting. *Neuron* 68, 815–834. doi: 10.1016/j.neuron.2010.11.022
- Calder, A. J., Lawrence, A. D., and Young, A. W. (2001). Neuropsychology of fear and loathing. *Nat. Rev. Neurosci.* 2, 352–363. doi: 10.1038/35072584
- Dayan, P., and Balleine, B. W. (2002). Reward, motivation and reinforcement learning. *Neuron* 36, 285–298. doi: 10.1016/S0896-6273(02)00963-7
- Dumas, G., Nadel, J., Soussignan, R., Martinier, J., and Garner, L. (2010). Inter-brain synchronization during social interaction. *PLoS ONE* 5:e12166. doi: 10.1371/journal.pone.0012166
- Field, T. M. (1981). Infant gaze aversion and heart rate during face-to-face interactions. *Infant Behav. Dev.* 4, 307–315. doi: 10.1016/S0163-6383(81)80032-X
- Itoha, K., Miwab, H., Takanobud, H., and Takanishi, A. (2005). Application of neural network to humanoid robots—development of co-associative memory model. *Neural Netw.* 18, 666–673. doi: 10.1016/j.neunet.2005.06.021
- Kakade, S., and Dayan, P. (2002). Dopamine: generalization and bonuses. *Neural Netw.* 15, 549–559. doi: 10.1016/S0893-6080(02)00048-5
- Kaplan, F., and Oudeyer, P.-Y. (2007). In search of the neural circuits of intrinsic motivation. *Front. Neurosci.* 1, 225–236. doi: 10.3389/neuro.01.1.1.017.2007
- Koch, P., Konen, W., and Hein, K. (2010). “Gesture recognition on few training data using slow feature analysis and parametric bootstrap,” in *2010 International Joint Conference on Neural Networks* (Barcelona). doi: 10.1109/IJCNN.2010.5596842
- Nagai, Y., Hosoda, K., Morita, A., and Asada, M. (2003). A constructive model for the development of joint attention. *Connect. Sci.* 15, 211–229. doi: 10.1080/09540909310001655101
- Ogino, M., Ooide, T., Watanabe, A., and Asada, M. (2007). “Acquiring peekaboo communication: early communication model based on reward prediction,” in *Proceedings of the IEEE 6th International Conference on Development and Learning* (London), 116–121. doi: 10.1109/DEVLRN.2007.4354053
- Oudeyer, P.-Y., Kaplan, F., and Hafner, V. V. (2007). Intrinsic motivation systems for autonomous mental development. *IEEE Trans. Evol. Comput.* 11, 265–286. doi: 10.1109/TEVC.2006.890271
- Phillips, G. D., Salussolia, E., and Hitchcott, P. K. (2010). Role of the mesoamygdaloid dopamine projection in emotional learning. *Psychopharmacology* 210, 303–316. doi: 10.1007/s00213-010-1813-z
- Reis, H. T., Sheldon, K. M., Gable, S. L., Roscoe, J., and Ryan, R. M. (2000). Daily well-being: the role of autonomy, competence, and relatedness. *Pers. Soc. Psychol. Bull.* 26, 419–435. doi: 10.1177/0146167200266002
- Russell, J. A. (1980). A circumplex model of affect. *J. Pers. Soc. Psychol.* 39, 1161–1178. doi: 10.1037/h007714
- Ryan, R. M., and Deci, E. L. (2000). Intrinsic and extrinsic motivations: classic definitions and new directions. *Contemp. Educ. Psychol.* 25, 54–67. doi: 10.1006/ceps.1999.1020
- Schultz, W. (1998). Predictive reward signal of dopamine neurons. *J. Neurophysiol.* 80, 1–27.
- Singer, T., Seymour, B., O'Doherty, J., Kaube, H., Dolan, R. J., and Frith, C. D. (2004). Empathy for pain involves the affective but not sensory components of pain. *Science* 303, 1157–1162. doi: 10.1126/science.1093535
- Stern, D. N. (1985). *The Interpersonal World of Infant: A View from*

- Psychoanalysis and Developmental Psychology.* New York, NY: Basic Books.
- Striano, T. (2004). Direction of regard and the still-face effect in the first year: does intention matter? *Child Dev.* 75, 468–479. doi: 10.1111/j.1467-8624.2004.00687.x
- Triesch, J., Teuscher, C., Deák, G. O., and Carlson, E. (2006). Gaze following: why (not) learn it? *Dev. Sci.* 9, 125–147. doi: 10.1111/j.1467-7687.2006.00470.x
- Watanabe, A., Ogino, M., and Asada, M. (2007). Mapping facial expression to internal states based on intuitive parenting. *J. Rob. Mech.* 19, 315–323.
- White, R. W. (1959). Motivation reconsidered: the concept of competence. *Psychol. Rev.* 66, 297–333. doi: 10.1037/h0040934
- Wiskott, L., and Sejnowski, T. J. (2002). Slow feature analysis: unsupervised learning of invariances. *Neural Comput.* 14, 715–770. doi: 10.1162/089976602317318938
- Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.
- Received:* 21 May 2013; *paper pending published:* 22 July 2013; *accepted:* 22 August 2013; *published online:* 12 September 2013.
- Citation:* Ogino M, Nishikawa A and Asada M (2013) A motivation model for interaction between parent and child based on the need for relatedness. *Front. Psychol.* 4:618. doi: 10.3389/fpsyg.2013.00618

This article was submitted to Cognitive Science, a section of the journal Frontiers in Psychology.

Copyright © 2013 Ogino, Nishikawa and Asada. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



From self-assessment to frustration, a small step toward autonomy in robotic navigation

Adrien Jauffret^{1*}, Nicolas Cuperlier¹, Philippe Tarroux² and Philippe Gaussier¹

¹ Neurocybernetic Team, Equipes Traitement de l'Information et Systèmes Laboratory, UMR 8051, Cergy, France

² Cognition Perception et Usages Team, Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur Laboratory, CNRS UPR 3251, Orsay, France

Edited by:

Marco Mirolli, Istituto di Scienze e Tecnologie della Cognizione, Italy

Reviewed by:

Frédéric Alexandre, INRIA Bordeaux Sud-Ouest/LaBRI-CNRS/IMN-CNRS/Université de Bordeaux, France

Vieri G. Santucci, Istituto di Scienze e Tecnologie della Cognizione - Consiglio Nazionale delle Ricerche, Italy

***Correspondence:**

Adrien Jauffret, Neurocybernetic Team, Equipes Traitement de l'Information et Systèmes Laboratory, UMR 8051, 2 Avenue Adolphe Chauvin, Cergy-Pontoise 95302, France
e-mail: adrien.jauffret@ensea.fr

Autonomy and self-improvement capabilities are still challenging in the fields of robotics and machine learning. Allowing a robot to autonomously navigate in wide and unknown environments not only requires a repertoire of robust strategies to cope with miscellaneous situations, but also needs mechanisms of self-assessment for guiding learning and for monitoring strategies. Monitoring strategies requires feedbacks on the behavior's quality, from a given fitness system in order to take correct decisions. In this work, we focus on how a second-order controller can be used to (1) manage behaviors according to the situation and (2) seek for human interactions to improve skills. Following an incremental and constructivist approach, we present a generic neural architecture, based on an on-line novelty detection algorithm that may be able to self-evaluate any sensory-motor strategies. This architecture learns contingencies between sensations and actions, giving the expected sensation from the previous perception. Prediction error, coming from surprising events, provides a measure of the quality of the underlying sensory-motor contingencies. We show how a simple second-order controller (emotional system) based on the prediction progress allows the system to regulate its behavior to solve complex navigation tasks and also succeeds in asking for help if it detects dead-lock situations. We propose that this model could be a key structure toward self-assessment and autonomy. We made several experiments that can account for such properties for two different strategies (road following and place cells based navigation) in different situations.

Keywords: bio-inspired robotics, self-assessment, action selection, metalearning, sensory-motor system, neural-networks

1. INTRODUCTION

Autonomy, in the field of robotics, is still an open and poorly defined problem for which concepts remain to be invented. By autonomous, we mean a system able to develop and evaluate their skills and decide whether its behavior is relevant or not according to the context. When we talk about autonomy relative to the behavior, we mean the ability to learn behaviors in an open-ended manner but also to manage them. The concepts of open-ended development and cumulative learning have been studied for years in psychology, machine learning and robotics. Those capacities highly depend on intrinsic motivations, involved in exploration and curiosity. The study of intrinsic motivations is gaining more and more attention lately as artificial systems face autonomous cumulative learning problems. Several computational models have been proposed to overcome these problems where some are based on the knowledge of the learning system, while others are based on its competence. The first knowledge-based model, proposed by Schmidhuber (1991), consisted in a world model that learned to predict the next perception given the current one and the action. Prediction progress is used as an intrinsic reward for the system. More recently, a similar model of artificial curiosity proposed by Oudeyer et al. (2007) allows an agent to focus on novel stimuli to improve its learning in

challenging situations, avoiding well known and totally unknown ones. The first competence-based model was proposed by Barto et al. (2004) where the intrinsic reward is given on the basis of the inability of the system to reach its goal. In the same way, Baldassarre and Mirolli (Schembri et al., 2007; Santucci et al., 2012) proposed a reinforcement learning architecture that implements skills on the basis of experts. See (Mirolli and Baldassarre, 2013) for a global state of the art of both approaches.

Here, we do not address the problem of learning new skills in a fully autonomous and open-ended manner. We propose in a first step to study open-ended development through the framework of human interaction. In our view, the agent requires a teacher to learn from demonstration but not in a prescriptive way. Since a long time, we develop models allowing robots to learn autonomously different navigation tasks such as: using latent learning to build a cognitive map to be able to reach several goals (Gaussier et al., 2002; Giovannangeli et al., 2006; Laroque et al., 2010; Cuperlier et al., 2007; Hasson et al., 2011; Hirel et al., 2013) or to use explicit or implicit reward to learn different kind of sensori-motor behaviors (Andry et al., 2002, 2004). Yet, it cannot be avoided that at some point the robot fails because of an incomplete learning or because of some changes in the environment. For complex task learning, autonomy means also being able to ask for

help and/or to learn from others. It can be made through several ways from stimulus enhancement, response facilitation to different level of imitation. Here we propose to focus on how to allow a system that can be fully autonomous (see previous papers) to work for some time under the supervision of a trainer in order to discover efficiently the solution of a complex problem and/or to complete its learning. To fit natural low level interactions, the training is performed thanks to a leash. Like a dog or a horse, our robot cannot avoid turning its head in the direction of the force applied on the leash. One key difference with a classically supervised system is that the robot has to evaluate when and how to take into account the supervision signal. For sake of simplicity, we will suppose, in this paper, there is no opposition between the robot needs and the trainer requirement. Hence we will not focus on how the robot use latent learning for building some cognitive map or exploit various reinforcement signals to modify its sensory-motor associations [see (Hirel et al., 2013) for the use of a similar architecture focusing on these issues].

Adding self-assessment capabilities to robots should be an interesting solution in this framework of social robotics where humans play a role in the cognitive development of the robot. It could allow the robot to seek for more interactions when needed as in the collaborative control system of Fong et al. (2003). The robot could also communicate its inability to improve its learning. It means that not only does the robot need to code its knowledge but also the limits of its knowledge. This becomes all the more important in integrated robotic systems which have to make decisions based on observations drawn from multiple modalities (Zillich et al., 2011).

Evaluating behavior performance requires the ability to predict the behavior itself at first. Then, it requires to detect potential problems by considering aspects of novelty in these predictions. Novelty is thus an important signal to consider since it represent a key feature providing feedbacks on the behavior's quality. The problem of self-assessment is then sensibly close to the class of novelty detection problems. Novelty detection is a commonly used technique to detect that an input differs in some respect from previous inputs. It is a useful ability for animals to recognize an unexpected perception that could be a potential predator or a possible prey. One of the main goal for self-assessment is self-protection. It is strongly used in situations that caused a failure or a threat in the past, or in the prediction of a threat or a future challenge (Taylor et al., 1995). It reduces the large amount of information received by the animal so that it can focus on unusual stimuli [see (Marsland, 2002) for a global state of the art].

A variety of novelty filters has been proposed where most of them work by learning a representation of a training set (containing only normal data), then trying to underline data that differ significantly from this training set. In the literature, one can find different classes of methods such as statistical outlier detection, novelty detection with supervised neural networks, techniques based on self-organizing map and gated dipole methods.

The standard approach to the problem of outlier detection (Sidak et al., 1967; Devroye and Wise, 1980) is to estimate the unknown distribution μ of a set of n independent random variables in order to be able to detect that a new input X does not belong to the support of μ . In the same way, extreme value

theory (Gumbel, 1958) focuses on distributions of data that have abnormal values in the tails of the distribution that generates the data.

The first known adaptive novelty filter is that of Kohonen and Oja (1976). It proposes a pattern matching algorithm where new inputs are compared with the best-matching learned pattern, meaning that non-zero output is only seen for novel stimuli. Self-organizing networks also provide solutions to detect novelty using unsupervised learning (Kaski et al., 1998) and particularly the so-called Adaptive Resonance Theory (ART) (Carpenter and Grossberg) network that uses a fixed vigilance threshold to add new nodes whenever none of the current categories represents the data. In a sense, the process of the ART network is a form of novelty detection depending on a vigilance threshold.

Supervised neural networks methods propose also novelty detection solutions by recognizing inputs that the classifier cannot categorize reliably. Such methods estimate kernel densities to compute novelty detection in the Bayesian formalism (Bishop, 1994; Roberts and Tarassenko, 1994).

Another solution is given by gated dipole fields, first proposed by Grossberg (1972a,b), then used to compare stimuli and model animal's attention to novelty (Levine and Prueitt, 1992).

Neural models of memory can also detect novelty by learning sequences of states that provide a simple mean of representing pathways through the environment (Hasselmo and McClelland, 1999). Dollé (2011) and Caluwaerts et al. (2012) propose models of metacontroller for spatial navigation that select on the fly the best strategy in a given situation. A competition following by a reinforcement learning allows to associate the action that best fits to the situation. Categorizing contexts are then required to recall the learned action.

Some studies propose that the hippocampus structure, besides its implication in spatial navigation, could be involved in novelty detection, since identifying novelty implies storing memories of normal situations and building expectations from these situations (Knight, 1996; Lisman and Otmakhova, 2001). The importance of novelty in emotional processes was suggested by appraisal theory (Lazarus, 1991; Scherer, 1984; Lewis, 2005). Novelty is closely related to surprise (which could be either positive or negative) but is also determinant in assessment processes for several other emotions (Grandjean and Peters, 2011). Emotional processes are particularly important in decision making while they can guide or bias behavior faster than rational processes, or when rational inferences are insufficient (Damasio, 2003). Griffiths proposes a taxonomy of emotions divided in 2 classes: primary emotions managed by the amygdala and cognitive ones that operate in the prefrontal cortex (Griffiths, 1997).

The role of emotions in communication are also important and have been studied in infants-adults interaction (Tronick, 1989). Infants show self-appraisal capability very early (before the age of 2) while trying to perform a task, but they show a few interest for parents approbation and focus on another goal in case of failure. From the age of 2 the children show reactions (crying, hooking on parents) when facing negative assessment (Stipek et al., 1992; Kelley et al., 2000).

Following the concept of bio-inspired robotics and a constructivist approach, we present integrated robotic control architecture

resulting from a close feedback loop between experiments on animals and robots. This leads to a better understanding of the mechanisms by which the brain processes spatial information. In the next sections we first present our previous model of visual place cells that allows a robot to exhibit simple and robust navigation behaviors (Gaussier et al., 2002; Banquet et al., 2005). Since this strategy has been successfully tested in small environments, we met issues while trying to navigate in larger and more complex ones (see Section 2.1.1). We propose to add a second strategy based on a simple, efficient and biologically plausible road following algorithm in order to overcome issues we met with the first one (see Section 2.1.2). Then, we propose a generic neural architecture able to evaluate both sensory-motor navigation strategies (see Section 2.2) based on novelty detection techniques. Finally, we show how a second-order controller, based on the computational literature on intrinsic motivations, can monitor novelty tendencies to modulate both strategies depending on their relevance in a given situation (see Section 2.3). We show how such a controller could also communicate the inability for the robot to perform its task, in order to learn from teacher demonstration. We claim that frustration could be a key feature to improve autonomy in an open-ended manner.

2. MATERIALS AND METHODS

2.1. TWO SENSORY-MOTOR NAVIGATION STRATEGIES

Here, we assume that a robot is given a repertoire of behaviors by the designer. In the following, we shortly present 2 of these behaviors that are available to the robot and on which evaluation and regulation mechanisms have been tested.

2.1.1. A model of Place Cells to perform sensory-motor navigation

In previous works, we developed a biologically plausible model of the hippocampus and entorhinal cortex in order to obtain visual place cells (VPCs) (O'Keefe and Nadel, 1978). VPCs are pyramidal neurons exhibiting high firing rates at a particular location in the environment (place field). Our model allowed controlling mobile robots for visual navigation tasks (Gaussier et al., 2002; Banquet et al., 2005).

1. A visual place cell (VPC) learns to recognize a constellation of landmarks-azimuths pattern in the panorama (see

Figure 1). VPCs activity depends on the recognition level of corresponding constellations. A winner-takes-all (WTA) competition selects the winning VPC [see (Giovannangeli et al., 2006) for more details].

2. Next, a neural network learns to associate a particular VPC with an action (a direction to follow in our case). The robot performs the action associated with the winning VPC. This sensory-motor architecture [Per-Ac (Gaussier and Zrehen, 1995)] allows the system to learn robust behaviors.

VPCs activity, even in outdoor conditions, shows a peak for the learned locations (see **Figure 2**) and generalizes quite correctly over large distances (2–3 m inside and 20–30 m outside).

Our architecture has been successfully tested in small sized environments (typically one room). However, our visual-only based mechanism shows limitations when trying to scale to larger and more complex environments (multi-room, outdoors). We encounter some situations where the large number of trees all around the system does not leave enough available landmarks to recognize a specific place (the entire panorama is full of green leaves that only represent noise for the system) and the only way to overcome such a problem is to follow the road below. We propose to overcome this issue by adding to our current architecture a biologically plausible road following strategy. Such a strategy allows the robot to follow roads rather than learning Place Cells, in situations where it is neither necessary, nor efficient to do so. Providing two different strategies to the robot is not sufficient by itself to navigate autonomously. The system also needs an action selection mechanism that evaluates both strategies (on the basis of a “meta” learning) to be able to select the right one in a given situation.

2.1.2. A model to perform road following behavior

In previous works, we presented a fast and robust biologically plausible road following strategy (Jauffret et al., 2013). Our algorithm consists in finding the best vanishing point among N potential points in an image. For example, let's consider 5 vanishing points equally distributed on the skyline. The robot will orient itself toward the winning vanishing point.

1. The system processes to an edge extraction of incoming images.

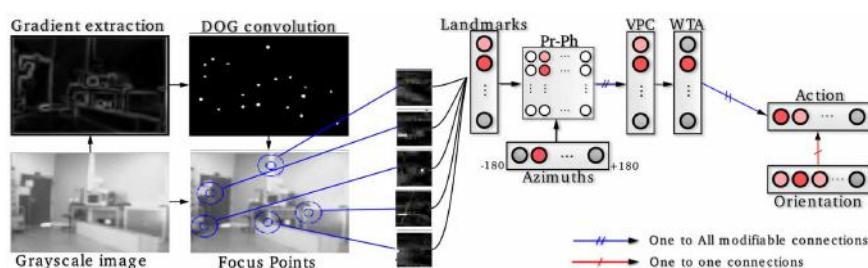


FIGURE 1 | Sensory-motor model of visual place cells (VPCs).

Gradient images are convolved with a DoG filter that highlights points of interest on which the system focuses on to extract local views. A VPC learns a specific landmarks-azimuths pattern. A winner-takes-all

competition (WTA) allows to select the winning VPC. Then, an association between the current action (robot's direction) and the winning VPC is learned, after what the system is able to move in such a direction each time that VPC wins.

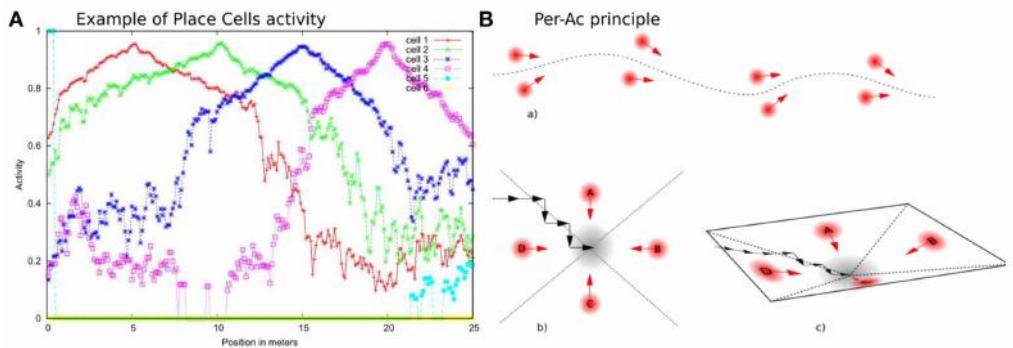


FIGURE 2 | (A) Activity of 4 PCs recorded on a linear track in a real outdoor environment. Each maximum of activity corresponds to the learned position of the corresponding PC. Our architecture provides good generalization properties since activities present large place field.

(B) PerAc principle: Only a few Place/Action associations are needed to perform simple and robust behaviors such as Path learning **(a)** or Homing **(b)**. The agent converges to the goal by falling into an attraction field **(c)**.

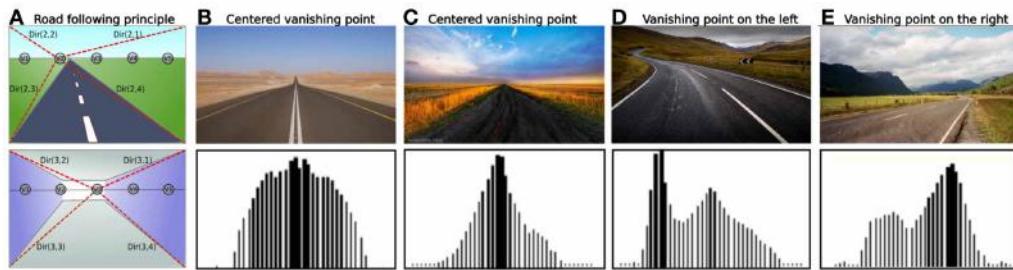


FIGURE 3 | (A) Road following principle for 5 vanishing point neurons (from V1 to V5) for outdoor and indoor cases (red dotted line: preferred directions of the winning neuron). Up: V2 is the best candidate. Down: V3 is the best candidate. **(B–E)** Results obtained with our algorithm on real images. For each cases: Up: real image of a road. Down: activity levels of 41 neurons, each one firing for a particular vanishing point location on the image. **(B)** Road with

boundaries: the vanishing point is well detected in the center of the image and generalized quite correctly to neighborhood **(C)** Without boundaries: the vanishing point remain salient since there is a significant gradient between road and grass. **(D)** Twisting road: 2 vanishing points are detected, one on the left and one in the middle. Nevertheless, the more active is the one on the left. **(E)** A vanishing point is detected on the right side.

- For each vanishing point considered corresponds one “vanishing” neuron V_n that integrates pixels whose edge orientation is aligned to this vanishing point (see Figure 3). The most active vanishing neuron corresponds to that where edges are mostly convergent to.
- Then, a simple WTA competition selects the best candidate between the N vanishing neurons.

The motor control of our model is directly inspired by control theories of Braitenberg vehicles (Braitenberg, 1986). This control is quite simple: when a vanishing point is detected on the right (resp. left), the robot will turn right (resp. left). Convergent behavior emerges from sensory-motor interactions between the system and its environment, without any need for an internal representation of the environment, or inference. Consequently, angular precision is less important than sample rate in such a control. We tested this algorithm on real images of road (see Figure 3) in several situations.

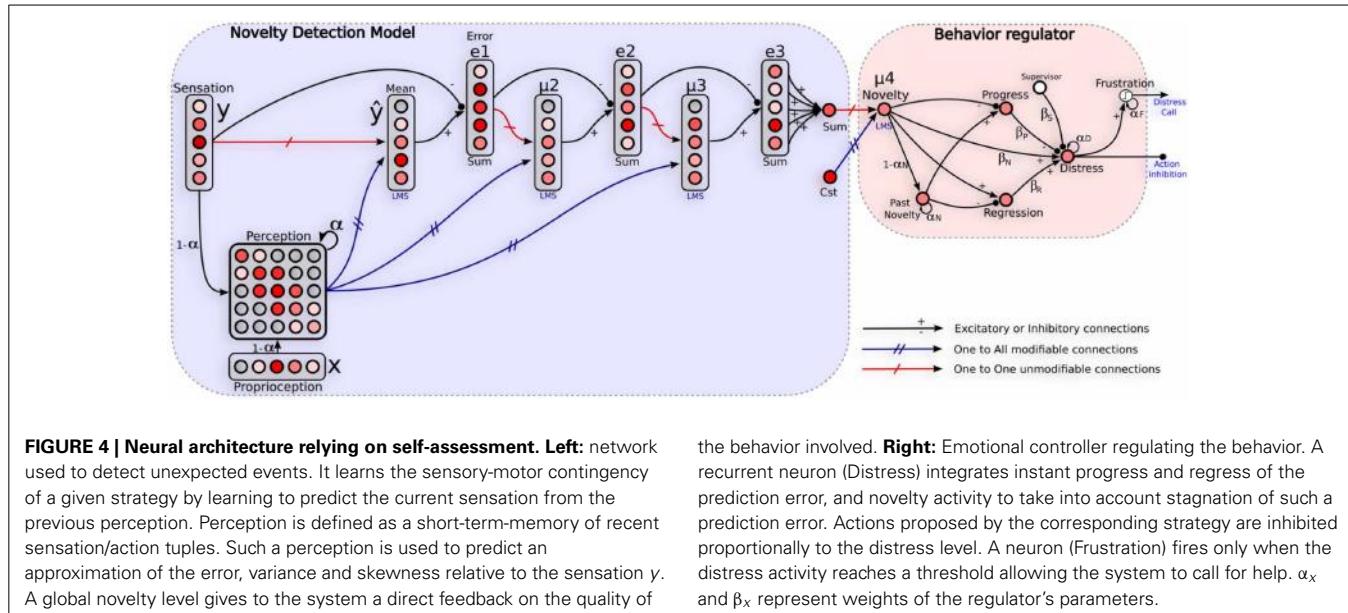
Our system succeeds in following any types of vanishing points such as roads, corridors, paths or railways. Furthermore, this algorithm has a satisfying framerate of 20 images per seconds

(for 41 vanishing point neurons tested on a I7 core processor) and this framerate increases when considering less neurons. Such a high framerate is obtained because only the higher gradients are considered in our algorithm. Therefore, the intensity of the gradient have been normalized by using the cosinus of the angle. So, in Figure 3 (case B) the gradient of road edges is not really high even though the vanishing point is detected.

A drawback of this method is the adjustment of the skyline position. Moreover, in some environments, the vanishing lines above the horizon can be an information (like in a corridor or in forest) although high reliefs or clouds can disturb localization of the vanishing point.

2.2. A NEURAL MODEL FOR NOVELTY DETECTION

Here we present a generic model for self-assessment based on novelty detection techniques. Our model consists in two steps. First, the learning of the sensory-motor contingencies induced by the navigation strategy involved in a normal situation (training set), second the ability to detect extraneous sensory-motor patterns in novel situations (see Figure 4).



2.2.1. Modeling the dynamic interaction between the agent and the environment

Learning to predict the sensory-motor contingencies of a strategy can be viewed as finding invariants in the robot's perception. In visual perception (Gibson, 1979), an affordance can be defined as building or accessing to an invariant characterizing one particular sensory-motor behavior. Based from this statement, we consider perception as the result of the learning of sensation/action associations allowing a globally consistent behavior (see (Gaussier et al., 2004; Maillard et al., 2005) for a complete mathematical definition of perception).

Following this assumption, we defined robot's perception as the integral of all its affordances. An affordance referring to a particular sensations/actions state:

$$Per(t) = \int_{-\infty}^t Sen^T(t).Ac(t).dt \quad (1)$$

Where Sen denotes a vector of sensations (sensory input), and Ac a vector of actions (given by agent's proprioception).

Lets denote y like Sen_i , a vector of n neurons y_i relative to agent's own sensations, $i \in N$. It can be both place cells or vanishing point cells in our case. y can be viewed as a set of random variable y_i . x is a vector of neurons x_i relative to agent's proprioception, where the winning neuron code for the current orientation. A matrix Per estimates the robot's perception by integrating sensations y and actions x in a finite shifting temporal window defined by the recurrent weight α . Per is the tensorial product between x and y with recurrent connections of weight α . It codes a short term memory of the agent's perception, where $Per_{i,j}$ denotes the particular tuple of both x_i and y_j neurons:

$$Per(t + 1) = \alpha.Per(t) + (1 - \alpha).Sen^T(t).Ac(t) \quad (2)$$

$$Per(t) = \sum_{i=0}^t \alpha^{i-1}.(1 - \alpha).Sen^T(t).Ac(t) \quad (3)$$

Basically, it means that recent inputs have a higher weight in our process than older ones. This type of filter has been tested by Richefeu and Manzanera (2006) in a motion detection context. The parameter α is used in order to attach more importance to the near past than to the far past.

2.2.2. Detecting novelty by processing absolute differences between predicted and real sensation:

Following this internal model of the robot's perception, we defined a vector \hat{y} , same size as y , that estimates the mean $E[y]$ of the current sensation y from the perception matrix Per by an online least mean square algorithm (LMS) (Widrow and Hoff, 1960): As a classical conditionning (Pavlov, 1927) the vector \hat{y} modifies on the fly the weights of connections coming from the perception matrix (unconditionned stimulus US) in order to estimate the sensation vector y (conditionned stimulus CS). We make the assumption that y follows a Gaussian distribution required by least-squares. An absolute difference between y and \hat{y} defines the instant error vector e . In the same manner, a vector \hat{e} , estimates the first moment about the mean $\mu_2 = E[e]$, of the current error $e = y - \hat{y}$, from the perception matrix Per by an online LMS algorithm. The second-order error is defined as $e_2 = e - \mu_2$. The second moment about the mean is defined as $\mu_3 = E[|e - \mu_2|]$.

The third order error is defined as $e_3 = e_2 - \mu_3$. Novelty N is defined as the global third moment about the mean. N is a single neuron that integrates all e_3 neurons activities: $N = |\sum_{i=1}^n e_3|$. N is summarized by:

$$N = E[||(|y - E[y]| - E[||y - E[y]||]) - E[||y - E[y]||]] - E[||y - E[y]||] \quad (4)$$

$$\text{where } ||y|| = \sqrt{\sum_{i=0}^n y_i^2 (L^2 - \text{norm})} \text{ or } \sum_{i=0}^n |y_i^2| (L1 - \text{norm}) \quad (5)$$

N represents the prediction error of the network, that will be used to detect unexpected events. Here, we defined novelty as the third order moment about the mean for empirical reasons while it is a good trade-off between precision and latency. Here, the different moments μ_2 , μ_3 , and μ_4 represent respectively the pseudo-variance, the pseudo-skewness and the pseudo-kurtosis while their measure follows the L1-norm rather than the L^2 -norm. Our architecture is thus able to learn an internal model of the dynamical interactions the system has with the external world.

2.3. MODELING FRUSTRATION TO REGULATE BEHAVIORS AND IMPROVE LEARNING

We showed that our model for self-assessment is able to give feedbacks on the quality of the behavior of the strategy involved. However, the system was not using such a confidence feedback to regulate its behavior. Here, we propose to implement an emotional controller able to make use of the novelty level, coming from the prediction mechanism. We propose that only considering the absolute novelty level is not sufficient to take correct decisions and regulate behaviors. First, short perturbations (small obstacles, sensor disturbance, visual ambiguities or singular false recognitions) should not affect so much the robot's behavior. Most of the time, the good generalization properties of our sensory-motor strategies allow the robot to stay inside the "attraction field" of the learned behavior (see **Figure 2**) and thus perform its task correctly, even if unexpected events appear. Because it is more interesting to consider the evolution of such a novelty activity rather than its absolute level, the agent should integrate the novelty activity over time and monitor its evolution to be able to judge its own behavior. If the prediction error remains high or increases whatever the agent tries, then the behavior should be considered as inefficient. And if this inefficiency is lasting this means the agent is caught in a deadlock. Similar assumptions have been proposed by Schmidhuber (1991) in a model-building control system driven by curiosity. This model deals with both problems of (1) do not take into account parts of the environment which are inherently unpredictable and (2) try to solve easy tasks before focusing on harder tasks. The author proposes to learn to predict cumulative error changes rather than simply learning to predict errors.

Based on previous works (Hasson et al., 2011), we propose to compute the instantaneous progress $P(t) = N(t - \delta_t) - N(t)$ and the instantaneous regress $R(t) = N(t) - N(t - \delta_t)$ as the derivatives of the novelty level $N(t)$.

Lets define an analog potential of frustration as a recurrent neuron that integrates instant progress and regress (see **Figure 4**). It also integrates novelty activity $N(t)$ to take into account stagnation of such a prediction error. This potential of frustration is called the distress level $D(t)$ in the followings. Actions proposed by the corresponding strategy are inhibited proportionally to this level. Frustration is then defined as a binary decision $F(t)$:

$$F(t) = \begin{cases} 1 & \text{if } D(t) > T \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

with T a threshold parameter, and $D(t)$ the distress level defined as:

$$D(t) = \alpha_D D(t - \delta_t) + \beta_S S + \beta_R R(t) - \beta_P P(t) + \beta_N N(t) \quad (7)$$

with $S(t)$ a reward coming from the supervisor and α_D , β_S , β_P , β_N weights of each variable. The binary frustration neuron fires only when the distress activity reach a threshold T (0.9 in our experiments). It allows the system to stop and call for help in order to improve its learning in novel situations.

2.4. SELECTING AND MERGING STRATEGIES WITH A DYNAMIC NEURAL FIELD

In our architecture, both strategies (place/action associations and vanishing point following) and their respective metacontroller run in parallel as independent channels (see **Figure 5**). Each strategy provides an action (an orientation) in a separate field of 361 neurons. Each neuron of the field codes for a particular orientation. Each field of action is inhibited proportionally to the distress level of the corresponding strategy. Both fields are merged into a global Dynamic Neural Field providing solutions for action selection/merging rather than a strict competition (Amari, 1977). The neural fields properties have already been successfully tested to move robot arms by imitation using visual tracking of movement (Andry et al., 2004), or motor control for the navigation of mobile robots (Schoner et al., 1995; Quoy et al., 2003). Neural Fields can account for interesting properties such as action selection according to contextual inputs or persistence in more detailed models (Prescott et al., 2001; Guillot-Gurnett et al., 2002). The neural field equation proposed by Amari is the following:

$$\tau \cdot \frac{u(x, t)}{dt} = -u(x, t) + I(x, t) + h + \int_{z \in V_x} w(z) \cdot f(x - z, t) dz \quad (8)$$

where $u(x, t)$ refers to the activity of neuron x (coding for an angle), at time t . $I(x, t)$ is the input to the system. h is a negative constant that ensures the stability. τ is the relaxation rate of the system. w is the interaction kernel in the neural field activation. A difference of Gaussian (DOG) models these lateral interactions that can be excitatory or inhibitory. V_x is the lateral interaction interval defining the neighborhood. Properties of this equation allow the computation of attractors corresponding to fixed points of the dynamics and to local maxima of the neural field activity. Selecting or merging multiple actions depends on the distance between them. Indeed, if two inputs are spatially close, the dynamic gives rise to a single attractor corresponding to the average of them (merging). Otherwise, if we progressively amplify the distance between inputs, a bifurcation point appears for a critical distance. The previous attractor becomes a repeller and two new attractors emerge. Oscillations between multiple actions are avoided by the hysteresis property of this competition/cooperation mechanism. Finally, a simple derivative of the robot orientation allows for motor control of the robot speed [see (Cuperlier et al., 2005) for more details].

2.5. EXPERIMENTAL SETUP

We have tested our models in several situations for both strategies. We first present experiments running in simulation showing the model principles. Next, we present an experiment with a real robot showing how our model deals with known difficulties of real life experiments such as odometry correction, noisy sensors, dynamic of obstacles, people moving, lights changing, etc.

2.5.1. Simulations

We used a 40×40 cm wide simulated robotic platform (see **Figure 6A**) equipped with 2 wheels, proximity sensors for obstacle avoidance and a pan-tilt camera used to extract points of interest in the visual panorama and a fixed camera to perceive vanishing points (a copy of our robulab platform from Robosoft). Our simulation software (Webots from Cyberbotics) provides physically realistic model for the robot and obstacles but neither noise nor 3D objects near walls are taken into account (2D realistic snapshots of our lab are simply stuck on simulated walls).

Setup 1: The place/action strategy is put ON while the road following strategy stays OFF. The purpose of the experiment is to test the self-assessment mechanism on the place/action strategy.

The environment is a simulated room of 15×15 m (see **Figure 6B**) with a uniform floor and salient landmarks on walls. The robot is trained by a human teacher (supervised learning) to perform a round path by learning Place Cell/Action associations. No more than 8 place/action associations were sufficient for the robot to perform a robust round trip in our experiment. A second smaller room is unknown by the system as no places have been learned in it. Consequently, navigating in this room results in inconsistent movements. The evaluation mechanism learns to predict the sensory-motor contingencies of the place/action strategy while the robot performs its round trip in a normal situation (similar to the training set). In this setup, the vector of sensation *Sen* is defined by the vector of 8 Place Cells learned by the system. We set the recurrent weight $\alpha = 0.95$ empirically, based on the frequency of changes in sensations. The sensory-motor loop of that strategy is quite slow since states only change when the robot navigates from one place to another (it mainly depends on the distance between 2 places and the robot's linear velocity). Indeed, an α near to 1 results in a long temporal window (old states are more important than recent ones). 3 laps were necessary for the evaluation mechanism to completely predict its sensation from all sensory-motor states perceived during the trip. Indeed, learning

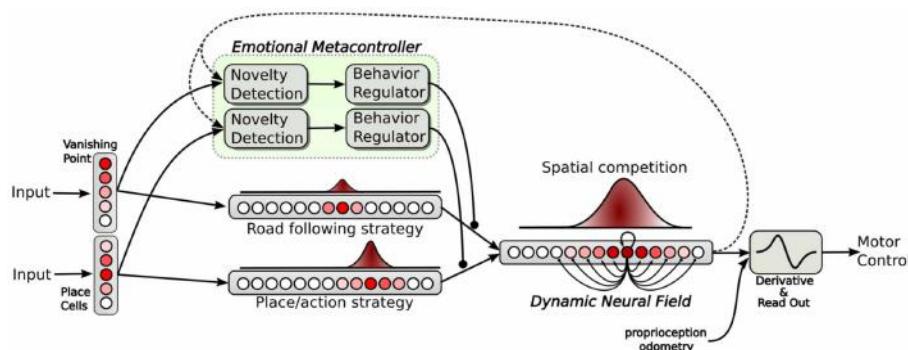


FIGURE 5 | Neural architecture relying on action selection. Down: both strategies provide an action in a field (each neuron of the field coding for a particular orientation). All action fields are merged into a dynamic neural field. This neural field provides a solution for decision making by selecting or merging actions in a robust manner (dynamic

attractors) and also provides good properties such as temporal filtering. Up: an emotional metacontroller learns to predict both strategies from its sensations and from the action proposed by the neural field (feedback link). Distress levels, depending on prediction errors, are used to modulate the action choice.

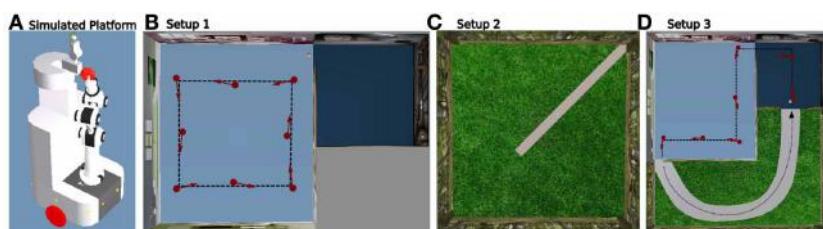


FIGURE 6 | Setup in Webots simulated environments. (A) Robotic platform used in our simulations. (B) Setup 1 used to test the place/action strategy. The system evolves in a simulated room of 15×15 m. It learns a few places associated with different actions (in red) to perform an ideal round behavior (black dotted line). (C) Setup 2 used to test the road following strategy. The system evolves in a simulated

outside environment of 40×40 m and can navigate both on road or grass. (D) Setup 3 used to test strategies selection/cooperation. A road links both rooms. No more than 7 place/action associations are needed for the robot to perform the entire loop (black dotted line). It does not need to learn anything off road while following the road is sufficient here.

is completed only when the novelty level reaches a minimum (typically below 0.4) and remains flat in all places.

Setup 2: The road following strategy is put ON while the place/action strategy stays OFF. The purpose of the experiment is to test the self-assessment mechanism on the road following strategy. The environment is a simulated garden of 40×40 m (see **Figure 6C**) with a white road passing over grass on the floor and trees texture on walls. The system is able to correctly follow roads when one is in its field of view. On the other side, navigating on grass results in random movements since there is no stable and well-defined vanishing point to follow. The evaluation mechanism learns to predict the sensory-motor contingencies of this strategy while the robot performs road following in a normal situation (training set). In this setup, the vector of sensation y is defined by 13 vanishing point neurons processed by the system. We set the recurrent weight $\alpha = 0.7$ empirically, based on the velocity of the sensory-motor loop. Indeed, the sensory-motor loop of that strategy is significantly faster than for the place/action strategy while vanishing point states change at a speed that directly depends on the robot's angular velocity. 2 min. of navigation were sufficient for the system to completely predict its sensation from sensory-motor situations perceived while following a road. Learning is completed when the novelty level reaches a minimum of 0.4 and stagnates.

Setup 3: This time, both strategies are active and run in parallel. The purpose of the experiment is to test strategies cooperation in a complex environment that is a mix of Setup 1 and 2 (see **Figure 6D**). A road is now linking both rooms by an outdoor part so that the robot can perform the entire loop. The system is trained to perform the loop: passing through both rooms and outside environment. Our model allows the system to correctly perform the entire loop by the learning of only 7 place/action associations. Indeed, the system does not need to learn any place on the outside part while following the road is sufficient in that part to perform the desired task. Consequently, the teacher does not have to correct the system in that part since the behavior resulting from the road following strategy is already the desired one.

2.5.2. Experiment on real robot

The following experiment runs in a real indoor environment (part of our laboratory). We used a real robotic platform similar to the simulated one (see **Figure 7A**). The environment is composed by 2 different rooms and a corridor (see **Figures 7C–E**). The place/action strategy is put ON while the road following strategy stays OFF. The purpose of the experiment is to test the frustration mechanism on the place/action strategy on a real robot experiment. The task for the robot is to achieve a complete loop passing through both rooms and corridor. 14 place/action were necessary for the robot to learn to perform the loop (see **Figure 7F**). As a stereotypical human/dog training interaction (Giovannangeli et al., 2006), the teacher uses a leash to pull the robot in the desired direction (see **Figure 7B**). Thus, the robot is detecting prediction error by comparing human order to its own will. This novelty detection neuromodulates the vigilance of the system so that it decides to recruit a new place cell and learns the association to its current orientation. Following such interactions, the robot is able to learn the path the human is teaching. A prescriptive learning (correcting the system rather than showing it the path) is necessary to get a stable and robust attraction field.

3. RESULTS

3.1. RESULTS RELYING ON NOVELTY DETECTION EXPERIMENTS

After the system has completely learned the desired trajectory (see Setup 1, **Figure 6**) and also learned to predict the sensory-motor contingencies relative to this trajectory, we tested it in several situations to show the ability for our model to detect whether such a situation is normal or abnormal.

In a first experiment, we tested the robustness of the strategy in a normal situation (see **Figure 8A**). The robot performs 12 standard laps without disturbance. Results show a robust and stable behavior with a trajectory close to the desired one. The novelty level stays relatively low since it never gets over 0.4, with a mean value of 0.2. It defines the minimum prediction error the system is able to achieve for this task. Such a minimum error is directly linked to the degree of deepness of the prediction process (the n th

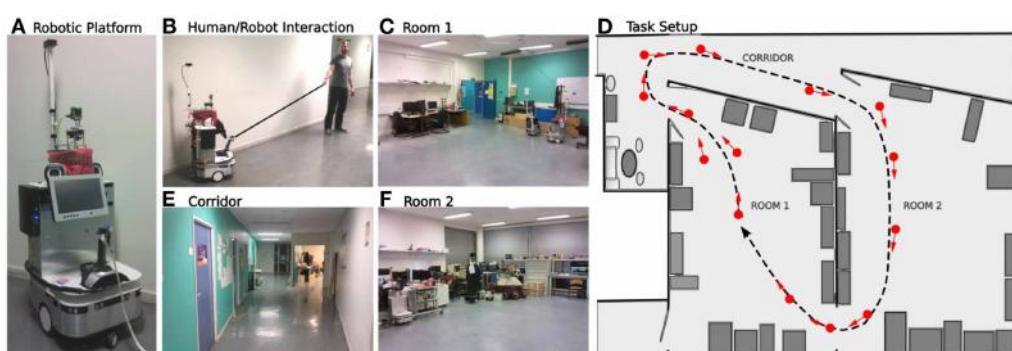
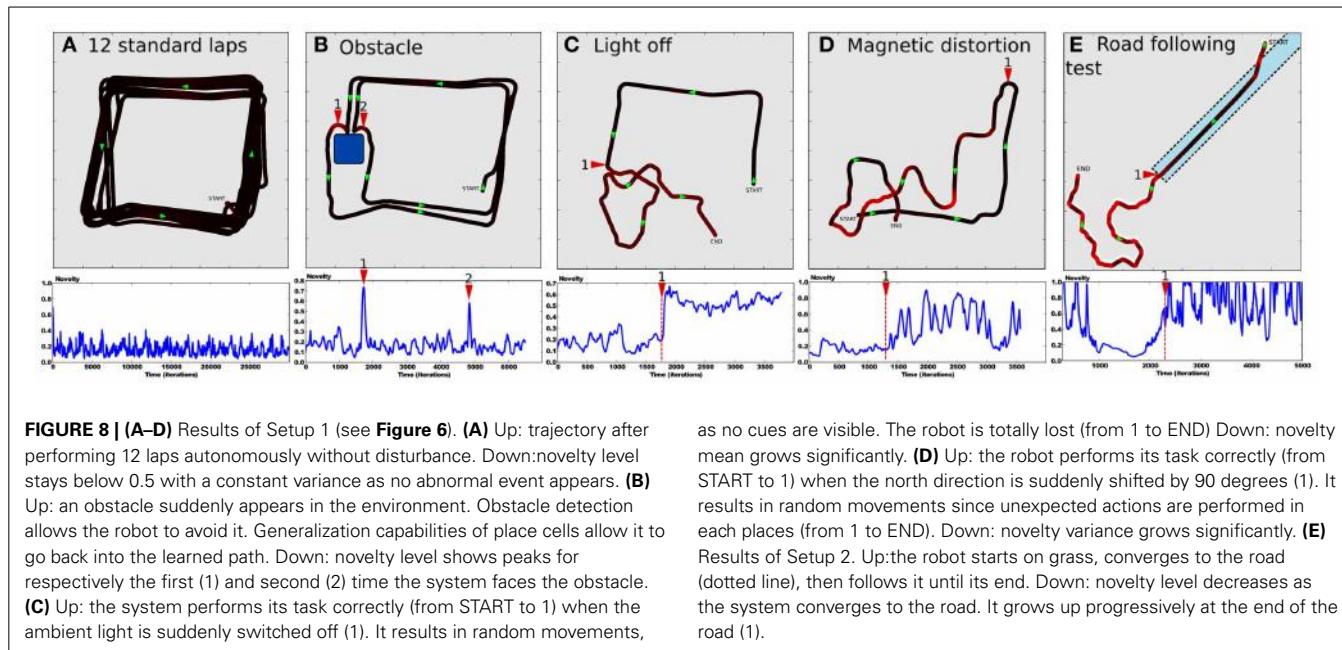


FIGURE 7 | Experimental setup in our laboratory. **(A)** Robotic platform used (Robulab from Robosoft). **(B)** Supervised learning: the teacher uses a leash to pull the robot in the desired direction. The robot learns the path autonomously. **(C–E)** The 3 different rooms of the experiment.

(F) Learned behavior. The robot learns to travel from room 1, passing through the corridor, to room 2, then back to room 1. About 14 place/action associations (red arrows) are learned to perform an ideal loop (black dotted line).



pseudo-moment about the mean). Since we defined the novelty as the third pseudo-moment about the mean, our model is not able to characterize statistical variations over such a precision. A fourth and fifth moment should be able to respectively learn the novelty mean and its variance.

In a second experiment, we introduce an obstacle in the environment so that the robot is forced to avoid it (see **Figure 8B**). Direct priority is given to the obstacle avoidance strategy by a subsumption architecture. The system avoids the obstacle, then successfully goes back to its original path thanks to the generalization properties of place cells/action associations. Novelty level shows peaks when the robot is avoiding the obstacle, since the orientation taken does not correspond to the learned one in that place.

In a fourth experiment, the light is suddenly put OFF while the robot performs its task (see **Figure 8C**). Consequently, the visual system is not able to maintain coherent place cells activity and the robot becomes totally lost. It results in random movements. Novelty level shows a sudden offset after the light is put OFF but keeps more or less the same variance. Indeed, the system is not able to recognize places anymore, even if it tries to predict it.

In the same way, a fifth experiment proposes to suddenly shift the north direction by 90 degrees (see **Figure 8D**). The system performs its task when the north is suddenly shifted. The robot behavior tends to be random a few seconds after the event. The novelty level shows large variations. Indeed, the system sometimes takes an unexpected orientation, sometimes a predicted one.

Finally, a sixth experiment proposes to test generalization capabilities of the novelty detection mechanism on the road following strategy (see Setup 2, **Figure 6**). The environment contains one road stopping in the middle, and grass elsewhere. The robot starts on grass, in a corner, oriented toward the road (see **Figure 8E**). Results show that the robot converges to the road in

order to be aligned with the road, then it correctly follows it until its end. Finally, it ends its trip by random movements onto grass after leaving the road, as no coherent vanishing point is perceived. Novelty level shows a progressive decrease while the robot converges to the road, then stays minimum and quite stable while following it. Novelty level increases progressively when leaving the road and stays high until the end of the experiment.

We also tested the robustness of the self-evaluation mechanism on a 1 h navigation experiment (not shown here).

3.2. RESULTS RELYING ON FRUSTRATION EXPERIMENT

In this experiment, we highlight the need for a frustration mechanism to request help in distress situations. The robot has learned to perform a squared loop in a room (see Setup 1, **Figure 6**). In a first period, the robot performs its task without disturbance (see **Figure 9A**). Results show a robust behavior with a stable trajectory close to the desired one. The distress level stays relatively low (below 0.4), with a mean below 0.3. It is the minimum prediction error the system is able to achieve for this task in a normal situation.

After some time, the teacher suddenly interferes with the robot to deviate its trajectory toward the second room (see **Figure 9B**). The robot tries to perform its task by taking the orientation associated with the winning place cell. The distress level increases while there is no consistency in the perceived sensory-motor sequence. After a while, the system stops for a distress call when the distress level reaches a frustration threshold (0.9 in our experiment). The teacher assists the robot by pulling it in the right direction to escape the small room (see **Figure 9C**). The teacher correction pushes the system to learn a new place/action association and the prediction mechanism to learn to predict this new situation. The robot successfully escapes the small room. The distress level decreases fast because the interaction with the teacher acts as an inhibitory signal in our emotional model. Once

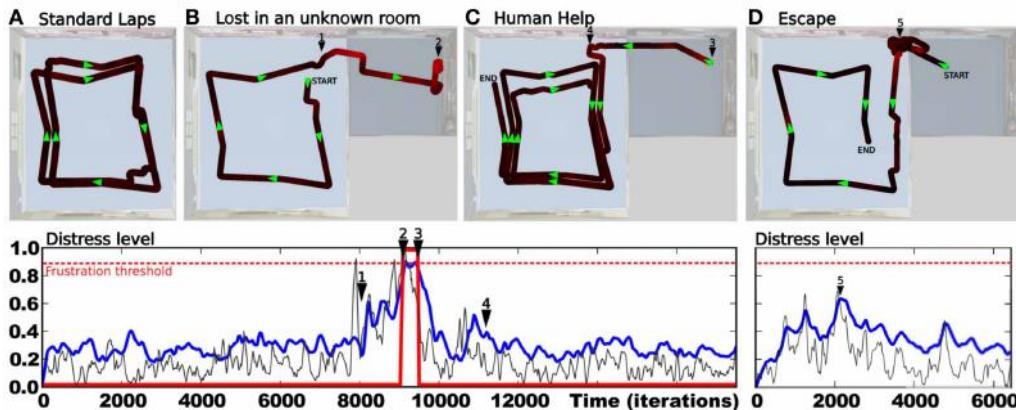


FIGURE 9 | Simulation highlighting the need for a frustration mechanism (see Setup 1, Figure 6). Up: (A) Trajectory after performing some laps autonomously without disturbance. (B) The supervisor suddenly interferes to deviate the robot into the second room (1). The robot tries to perform its task without success. The distress level increases progressively while there is no consistency in the perceived sensory-motor sequence. Finally, the system stops and call for help (2). (C) The supervisor assists the robot by pulling it in the right direction to escape this room (3). The system

learns the new place/action association . Once leaving the small room, it goes back to its stable attractor (4) and perform its task correctly. (D) After human demonstration, the simulated robot is able to escape autonomously from the small room. Down: evolution of the distress level in time. It increases as the robot becomes lost in the small room (1). It reaches a frustration threshold of 0.9 (red dotted line) (2), then goes back to normal after human help (4). It increases when the robot is between both rooms (5), indicating that a learning refinement is possible. However stays below the threshold.

leaving the small room, the robot goes back to the first room and converges again to a stable attractor. It continues performing its original task correctly. In another experiment, the robot starts in the small room, in a place different from the learned one (see Figure 9D). Since the robot already faces this situation in the past and thanks to the good generalization properties of place cells, it knows what to do to escape the room and to get back to its stable attractor. Results show that the robot takes the learned orientation to escape the room. The distress level stays low because the situation is considered as normal (predicted) this time. As the robot reaches the frontier between both rooms, its behavior tends to be a bit hesitating. This is due to place ambiguity since the robot hesitates between two place cells associated with contradictory actions. The distress level increases progressively. However, such an odd situation is not long enough to trigger a distress call, and the robot finally successes in getting back to its stable attractor. The distress level decreases slowly and the situation goes back to normal.

Such an interaction allows the system to learn from the teacher how to solve the problem so that it will be able to escape autonomously next time.

3.3. RESULTS RELYING ON STRATEGIES COOPERATION EXPERIMENT

In the following experiments, we highlight the need for an emotional controller to regulate behaviors to solve complex navigation tasks. Navigating in a wide and complex environment requires a metacontroller to make different strategies cooperate in a coherent manner. The robot has learned to perform a complete loop, passing from one room, to the other, to the garden, then back to the first room (see Setup 3, Figure 6). Both strategies and their evaluation mechanisms run in parallel. Strategies cooperate by proposing their actions weighted in real time by their own evaluation. Actions coming from the different strategies are merged

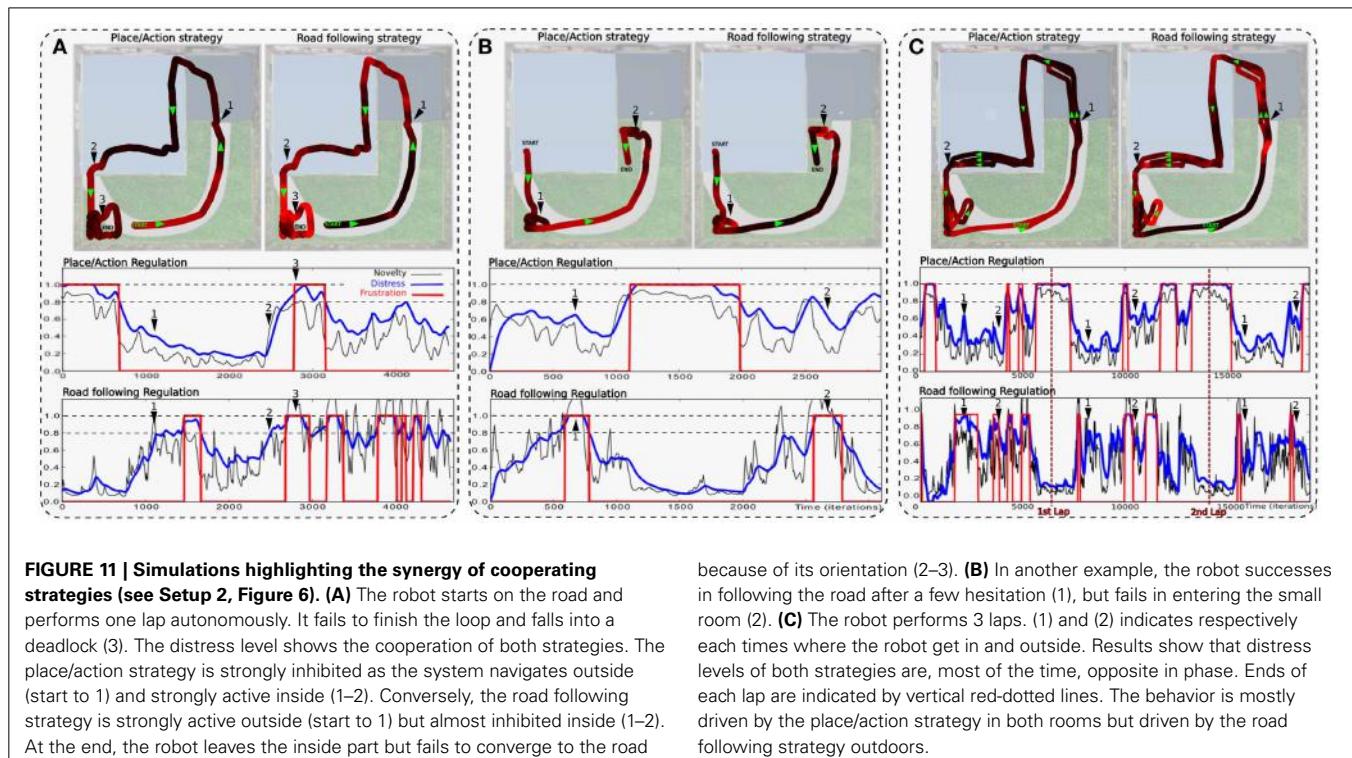
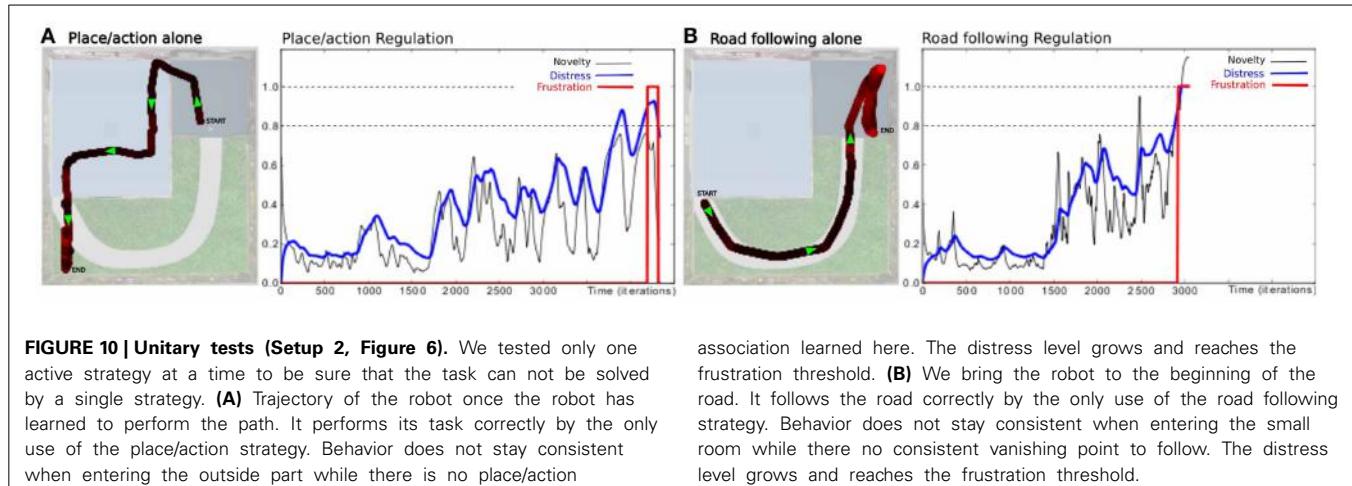
into a dynamic neural field that controls robot's movement (see Part.2.4). It allows a smooth cooperation rather than a strict competition. It is also important to notice here that the frustration mechanism does not trigger a distress call procedure in the following experiments. Since we want to test smooth cooperation, the robot does not stop even if both strategies reach the frustration threshold.

In a first simulation, we tested each strategy alone to ensure the system can not solve such a complex task with only a single strategy. When testing the place/action strategy (see Figure 10A), results show that the robot performs its task correctly in both room, but falls into a deadlock when navigating outdoors and finally get frustrated. In the same way, when testing the road following strategy (see Figure 10B), the robot follows the road correctly in the garden, but falls into a deadlock when entering a room and finally get frustrated.

In a second simulation, the robot performs one lap autonomously (see Figure 11A). It starts in the middle of the road at the bottom of the environment. It performs the loop correctly but fails to finish it and falls into a deadlock when entering the garden. It is due to the orientation the robot takes when leaving the room. If this orientation is too much different from the direction of the road, the system is not able to converge to it.

The distress level shows the cooperation of both strategies during the loop. The place/action strategy is strongly inhibited as the system navigates in the garden and strongly active inside. Conversely, the road following strategy is strongly active when navigating on road outside but almost inhibited inside.

In a third simulation, the robot starts in the first room, succeeds in converging to the road after a few hesitation, but fails to enter the small room (see Figure 11B). The reason for this success in converging to the road this time is mainly by chance. Next, the robot fails to enter the small room because the generalization



properties of place cells allow the robot to recognize this room before entering in it. As a result, it decided to turn too soon and falls into a local minimum. Such a problem can be solved by learning a new place/action association at the end of the road, ensuring to correctly enter the room.

In a fourth experiment, the robot performs 3 laps autonomously (see Figure 11C). The robot starts in the middle of the road at the bottom of the environment. Results show that, most of the time, the distress levels of both strategy are opposite in phase. The distress level of the place/action strategy stays low indoors as the sensory-motor sequence stays predictable. It is high outside while no discriminant landmarks are recognized and no places have been categorized in that

part. On the other hand, one can see that the distress level of the vanishing point following strategy is low when the robot follows the road outside. It is mostly high indoors while there is no consistent vanishing point to follow. Distress levels induce proportional inhibition of corresponding behaviors. Accordingly, the robot's behavior is mostly driven by the vanishing point following strategy while navigating on the road outside. Conversely, it is mostly driven by the place/action strategy on the inside part.

Beyond such predictable results, the experiment exhibits good properties that emerge from the synergy of both strategies. As a matter of fact, we encounter several situations where the cooperation enhance the performance obtained with a single

strategy. It usually happens in situations where a place/action association allows the robot to pass through a door. In several cases, the contrast induced by an open door make it be perceived as a coherent vanishing point by the system so that the robot naturally converges in its direction without the need for multiple and precise place/action associations. Despite the fact that such a property increases the quality of the behavior, it may be a constraint in others.

Finally, our results also underline some issues during transitions from a place/action to a road following strategy. Indeed, the teacher has to be careful pulling the robot in the direction of the road, otherwise the system can not evaluate the vanishing point correctly and allow the robot to follow the road. It is due to the delay the controller needs to evaluate a strategy. This can result in a deadlock situation where the system switches from one strategy to another without being frustrated enough to call for help.

3.4. RESULTS RELYING ON REAL ROBOT EXPERIMENT

The following experiments were performed in our laboratory using a robot similar to the simulated one (see **Figure 7**). The purpose of the experiment is to test the frustration mechanism on the place/action strategy on a real robot experiment (person and furniture moving, ambient light changing). The road following strategy is disabled. The robot has learned to perform a complete loop, navigating from the first room, to the corridor, to the second room, then back to the first one. The prediction mechanism starts to learn to predict the place/action contingencies after the system finishes the first lap (see **Figure 12**). We choose not to let the prediction mechanism learn during the first lap in order to get a stable behavior before the system tries to predict it.

The second lap corresponds to an intense metalearning stage since the predictor starts to learn and each place is new for the system (see **Figure 13A**). Distress level (Dl) shows peaks for each place. The Dl decreases while ending the loop, because the starting point is already predicted.

After this training stage, we let the robot performing its task for a while, correcting its trajectory only when needed (see **Figure 13B**). In this normal situation the distress level stays low (below 0.5) except for one area where it shows peaks. Such an area corresponds to a place where the robot navigates close to the window and is disturbed by the sunset at the time of the experiment. It means that the robot is less confident in its place/action strategy in this area. However it is not sufficient to frustrate the robot and the behavior stays coherent.

Later, the robot performs its task when the teacher suddenly deviates it into a small and unknown room (see **Figure 13C**). As the robot becomes lost, the distress level increases and finally reaches the frustration threshold. The robot stops and call for help. The teacher then assists the robot in escaping the room. The system learns that new state, gets out of the room and goes back to its stable attractor. The area where the robot were unconfident is now totally predicted since the sun is down and does not disturb it anymore.

4. DISCUSSION

In this paper, we have addressed two different roles of a self-assessment mechanism for long range and complex robotic

navigation. We presented its regulatory role in managing behaviors according to the situation, and its social role in communicating frustration to avoid deadlocks and improve learning.

First of all, we briefly presented our previous model of place cells that allows robots to perform simple and robust sensory-motor behaviors in small size environments. We highlighted the need to find solutions to overcome some issues we met while trying to navigate in more complex ones. We underlined situations where the number of available landmarks in the panorama is very low and the visual system deals with noisy information. We proposed to overcome these issues by taking into account other strategies. We extended our architecture by adding a robust and biologically plausible road following strategy that allows the robot to naturally converge to visible roads. Such a strategy allows to follow potential vanishing points instead of learning place/action associations, in situations where it is neither necessary, nor efficient to do so.

These behaviors defined the robot's skills for facing the situations encountered in the environment. However, these behaviors are in competition. The robot needs a second-order controller to manage them. Such a metacontroller needs a mechanism that evaluates behaviors. We argue that for evaluating its behavior, the system requires to monitor novelty in its predictions. Monitoring novelty or abnormality in the behavior is thus identified as a key feature for a second-order controller to manage robot's strategies. Following this statement, we proposed a model for self-assessment based on novelty detections in a dynamical point of view. In this view, the system must, at first, (1) learn to predict its sensations from its past perception in a training situation, next (2) monitor novelty and respond accordingly. We defined perception as an internal model of the sensory-motor interactions the system has with the external world. This model of perception provides a generic grounding to perform predictions on agent's sensation. The model could be adapted to any sensation/action loop and thus for a reasonable computational cost if one considers that, most of the time, sensation and action are correlated (except pathological cases). However, since the dimension of the "Perception" tensor might be large, it is important to define abstract input vectors (e.g. few landmark neurons instead of raw visual data) to avoid combinatorial explosion.

Even if we choose in this paper to stay at a theoretical level, the analogy with the computation that could be performed in the hippocampal system are strong enough to provide solutions to avoid scaling issues. In future works, we plan to replace the complete Per matrix by a sparse matrix where the encountered products could be learned by specific units (see our work on parahippocampal and perirhinal merging in prph (Giovannangeli et al., 2006)). To go one step further, the number of states could be also reduced if we avoid "grand mother" cells solution (both for landmarks and place cells) and replace them by a sparse coding allowing to use combinatorial aspects at our advantage. Another solution could be to represent the sensory activity by a compressed code having a low probability of ambiguity (something similar to a hash code or a random M to N projection with $M \gg N$ for instance).

It should also be noted that the model by itself could not learn different timing or periodic phenomena since

the recurrent weight α has to be set empirically for each sensory-motor behavior. But the problem should be solved if we consider a set of novelty detectors with different time constants.

Then, we proposed to estimate the different moments about the mean by an online least-mean-square algorithm. One can note that least-mean-square requires independent and identically distributed (iid) random examples to ensure its convergence.

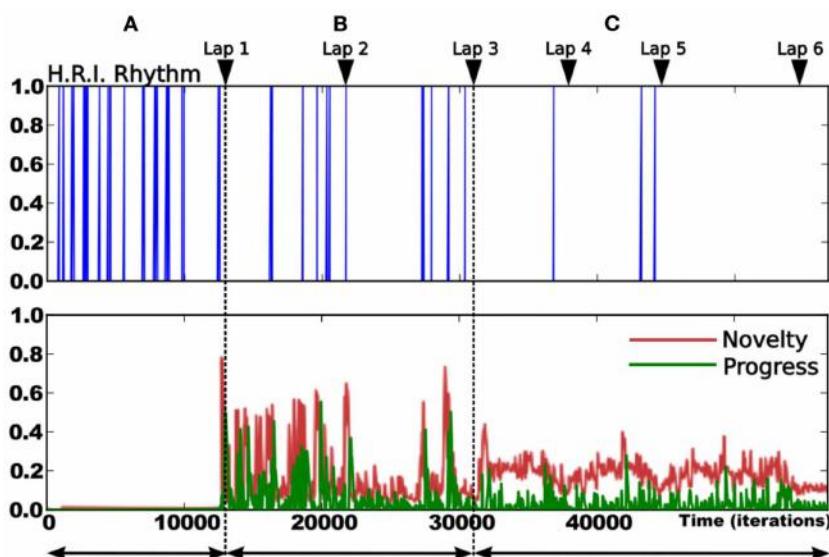


FIGURE 12 | Details of the learning stage over 6 laps. Up: rhythm of human/robot interactions (HRI): Dirac pulses correspond to guiding instants. The frequency of interaction decreases over time. It gives a direct measure of the system's autonomy (inversely proportional to frequency). Down: novelty (red) and Progress (green) level. We observe 3 different periods: **(A)** corresponds to the beginning of the learning session (first lap). The high frequency of pulses indicates that the human teacher is roughly directive as the robot does not know

anything about the task. The metacontroller is OFF during this stage. **(B)** corresponds to the evaluation stage (second and third lap). The teacher evaluates the robot's behavior and correct it only when needed. The metacontroller is ON and starts to learn to predict the sensory-motor sequences. Consequently, both novelty and progress levels are high during this stage. **(C)** corresponds to the final stage where the robot is autonomous enough to stop learning. Rare corrections are still needed at some points.

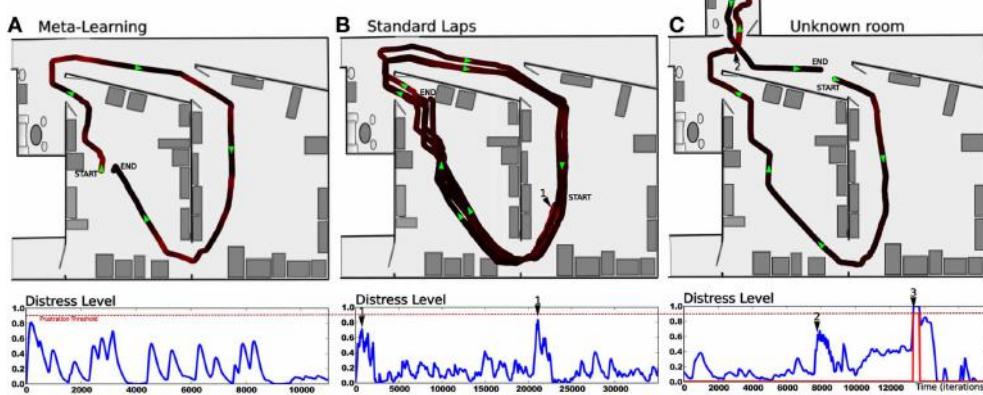


FIGURE 13 | Results relying on real robot experiment (see Figure 7). **(A)** 1 lap trajectory during the metalearning stage. The robot has already learned to perform the task and learns to predict it. The distress level is high each time the robot get from one predicted state to a unknown one. Then the distress level goes down as the robot learns to predict that unknown state. All sensory-motor states are almost predictable after one lap. **(B)** After the metalearning stage, we let the robot perform some laps autonomously. The distress level stays low

except in one particular place (1) which corresponds to a place where the robot is close to the window and is slightly disturbed by the sunset. The robot is less confident in this area, however it is not sufficient to frustrate it. **(C)** Later, the robot performs its task when the supervisor suddenly interferes to deviate the robot into an unknown room (2). It tries to perform its task without success and finally get frustrated (3). The supervisor shows how to escape the room, afterward the system is able to escape autonomously.

However, even if such iid examples are not available in our experiments, the iid constraint is negligible here. Indeed, thanks to the large dimension of the perception matrix, (1) the problem is linear (except pathological cases) and (2) the system does not care about any complex unlearning processes nor does it need a high precision on its output.

Novelty is then measured based on the deviation of the monitored perception from the expected one. It is defined as the prediction error at a n 'th level. Here, we defined novelty as the third order moment about the mean for empirical reasons. Actually, one can choose an arbitrary order to define novelty, depending on the desired accuracy. For example, modeling novelty by a first order error (simple difference between raw sensation and its average) results in a rather poor detection but decreases the time needed by the predictor to learn to predict the task. With such a poor system, periodical or sporadic events generates novelty as they differ from the average. Thus, they cannot be considered as normal by the system since variance is not taken into account. Conversely, defining novelty as a high order error results in a finer detection. However, in this case, the predictor needs a lot of time to completely predict a normal sensory-motor situation. Because of the online and memory-less constraints of our model, the estimation of a particular moment requires to wait for the estimation of each previous order. It raises few questions: Do animals predict in the same way? And do they have some estimation latency which increases by the level of precision?

Results showed that our novelty detection model presents good generalization capabilities since the same architecture can work at least for two different sensory-motor strategies.

Finally we showed that only considering the absolute novelty level was not sufficient to take correct decisions and regulate behaviors. For example, short perturbations might not affect so much the behavior. Most of the time, the robot stays in its attraction basin and performs the task correctly, even if unexpected events appear. The reason is that novelty does not refer by itself to a positive or negative reward. We showed that it is more interesting to consider the evolution of such a novelty activity rather than its absolute level. Monitoring the novelty tendency, by integrating its activity over time, provides a solution for the system to judge its own behavior. If the prediction error stagnates at a high level or if it increases whatever the robot tries, then the behavior should be considered as inefficient. And if this inefficiency is lasting this means the agent is caught in a deadlock.

Following these assumptions, we propose an emotional metacontroller (modeling frustration) that monitors prediction progress to modulate both strategies and adapt the behavior to the situation. We made several experiments that highlight the need for such a metacontroller to switch between strategies.

Moreover, we underline the role of emotions in communication by adding a simple distress call procedure triggered by the robot's frustration. This procedure allows the robot to communicate its inability to achieve the task by calling for help if no relevant strategies are found (if switching strategy does not increase any progress at all). Even if this procedure uses an ad-hoc distress call mechanism, it is triggered by a meaningful signal that point

out situations where a refinement is possible. However, our emotional controller is not sufficient by itself to reach a full autonomy. Unlike intrinsically motivated systems such as (Schmidhuber, 1991; Barto et al., 2004; Oudeyer et al., 2007; Schembri et al., 2007; Baranes, 2011; Santucci et al., 2012), our system still requires a teacher to learn from demonstration but not in a prescriptive way. In this paper, frustration is presented as a useful intrinsic motivation for the agent to gradually develop its autonomy in an open-ended but supervised manner. Future works will focus on how to make use of this internal signal to improve learning in a fully autonomous way (without the need for human supervision).

Yet, our model has 3 main drawbacks that we should solve in a near future:

- The recurrent weight α , that defines the short-term-memory of the agent's perception, has to be different from one strategy to an other. Indeed, it highly depends on the own dynamic of the strategy involved.
- The size of the sensation vector y has a direct impact on the prediction dynamics. Indeed, the impact of a sensation neuron y_i on the novelty level is divided by the number of neurons in y .
- Yet, the learning stage is still separated from the use stage. This leads to the first role of emotions that could allow the agent to directly regulate its learning. The system should then decide whether to learn a situation as normal or abnormal.

Current works focus on testing the model performance on long range outside experiments (navigating several kilometers) with a real outdoor robot. We also focus on how a simple feedback loop can help the system to disambiguate its perception in an active way. Because of the ambiguity of perception, our system sometimes needs changes in its sensation to be able to correctly measure the quality of a given strategy. Thus, we study how to use the prediction error as a feedback signal that modulates actions accordingly. A high prediction error will trigger a high noise on robot's actions, inducing changes in sensation. Such changes will decrease the prediction error only if sensory-motor contingencies become predictable, and will increase it if not. Behavioral alteration is directly proportional to the prediction error. We wish this homeostatic mechanism will allow the system to regulate itself, updating its knowledge by actively altering its behavior in order to check whether its expectation is true.

ACKNOWLEDGMENTS

The authors would like to thank M. Belkaid and C. Grand for their assistance in part of the modeling and experiments. Many thanks also to A. Pitti for the interesting discussions and guidance, A. Blanchard for the work he did on the Neural Network simulator used in these experiments and F. Demelo for the robots hardware support. Thanks to the AUTO-EVAL project, the NEUROBOT French ANR project and the Centre National de la Recherche Scientifique (CNRS) for the financial support.

REFERENCES

- Amari, S. (1977). Dynamics of pattern formation in lateral-inhibition type neural fields. *Biol. Cybern.* 27, 77–87. doi: 10.1007/BF00337259
- Andry, P., Gaussier, P., and Nadel, J. (2002). “From Visuo-Motor Development to Low-level Imitation,” in *Proceedings of the second workshop on Epigenetic Robotics*, (Lund: Lund University Cognitive Studies), 94.
- Andry, P., Gaussier, P., Nadel, J., and Hirsbrunner, B. (2004). Learning invariant sensory-motor behaviors: a developmental approach of imitation mechanisms. *Adapt. Behav.* 12, 117–140. doi: 10.1177/105971230401200203
- Banquet, J.-P., Gaussier, P., Quoy, M., Revel, A., and Burnod, Y. (2005). A hierarchy of associations in hippocampo-cortical systems: Cognitive maps and navigation strategies. *Neural Comput.* 17, 1339–1384. doi: 10.1162/0899766053630369
- Baranes, A. (2011). *Motivations Intrinseqes et Contraintes Maturationnelles pour l'Apprentissage Sensorimoteur*. PhThesis, D, INRIA - Sud Ouest: Université de Bordeaux.
- Barto, A., Singh, S., and Chentanez, N. (2004). “Intrinsically motivated learning of hierarchical collections of skills,” in *International Conference on Developmental Learning (ICDL)*, La Jolla.
- Bishop, C. M. (1994). Novelty detection and neural network validation. *IEEE Proc. Vis. Image Signal Process.* 141, 217–222. doi: 10.1049/ip-vis:19941330
- Braitenberg, V. (1986). *Vehicles: Experiments in Synthetic Psychology*. Cambridge: MIT Press.
- Carpenter, G. A., and Grossberg, S. (1988). The ART of adaptive pattern recognition by a self-organising neural network. *IComputer IEEE* 21, 77–88.
- Caluwaerts, K., Staffa, M., N'Guyen, S., Grand, C., Dollé, L., Favre-Félix, A., et al. (2012). A biologically inspired meta-control navigation system for the Psikharpx rat robot. *Bioinspir. Biomim.* 7:025009. doi: 10.1088/1748-3182/7/2/025009
- Cuperlier, N., Gaussier, Ph., Laroque, Ph., and Quoy, M. (2005). “Goal and motor action selection using an hippocampal and prefrontal model,” in *Model. Nat. Action Select.* (Edinburgh: AISB Press), 100–106.
- Cuperlier, N., Quoy, M., and Gaussier, Ph. (2007). Neurobiologically inspired mobile robot navigation and planning. *Front. Neurorobot.* 1:3. doi: 10.3389/neuro.12.003.2007
- Damasio, A. (2003). *Looking for Spinoza: Joy, Sorrow and the Feeling Brain*. San Diego, CA: Harcourt Inc.
- Devroye, L., and Wise, G. L. (1980). Detection of abnormal behavior via nonparametric estimation of the support. *Appl. Am. J. Math.* 38, 480–488. doi: 10.1137/0138038
- Dollé, L. (2011). *Contribution d'un Modèle Computational de Sélection de Stratégies de Navigation aux Hypothèses Relatives à l'apprentissage Spatial*. Paris: PhThesis, D, UPMC-Sorbonne université.
- Fong, T. W., Nourbakhsh, I., and Dautenhahn, K. (2003). A survey of socially interactive robots. *Robot. Auton. Syst.* 42, 143–166. doi: 10.1016/S0921-8890(02)00372-X
- Gaussier, P., and Zrehen, S. (1995). Perac: a neural architecture to control artificial animals. *Robot. Autonom. Syst.* 16, 291–320. doi: 10.1016/0921-8890(95)00052-6
- Gaussier, P., Revel, A., Banquet, J.-P., and Babeau, V. (2002). From view cells and place cells to cognitive map learning: processing stages of the hippocampal system. *Biol. Cybern.* 86, 15–28. doi: 10.1007/s004220100269
- Gaussier, P., Bacon, J. C., Prepin, K., Nadel, J., and Hafemeister, L. (2004). “Formalization of recognition, affordances and learning in isolated or interacting animats,” in *The Society for Adaptive Behavior SAB'04*, (Los Angeles, CA: MIT Press), 57–66.
- Giovannangeli, C., Gaussier, P., and Banquet, J. P. (2006). Robustness of visual place cells in dynamic indoor and outdoor environment. *Int. J. Adv. Robot. Syst.* 3, 115–124. doi: org/10.5772/5748
- Gibson, J. (1979). *The Ecological Approach to Visual Perception*. Boston, MA: Houghton Mifflin.
- Grandjean, D., and Peters, C. (2011). “Novelty processing and emotion: conceptual developments, empirical findings and virtual environments,” in *Emotion-Oriented Systems: The Humaine Handbook*, eds P. Petta, C. Pelachaud, R. Cowie (London: Springer), 441–458.
- Griffiths, P. E. (1997). *What Emotions Really Are: The Problem of Psychological Categories*. Chicago, IL: University of Chicago Press. doi: 10.7208/chicago/9780226308760.001.0001
- Grossberg, S. (1972a). A neural theory of punishment and avoidance. I. Qualitative theory. *Math. Biosci.* 15, 39–67. doi: 10.1016/0025-5564(72)90062-4
- Grossberg, S. (1972b). A neural theory of punishment and avoidance. II. Quantitative theory. *Math. Biosci.* 15, 253–285. doi: 10.1016/0025-5564(72)90038-7
- Guillot-Gurnett, A. K. N., Girard, B., Cuzin, V., and Prescott, T. J. (2002). “From animals to animats 7,” in *Proceedings of the Seventh International Conference on Simulation of Adaptive Behavior*, eds J. Hallam-Hayes G., B. Hallam, D. Floreano, and J. A., Mayer (Cambridge: MIT Press).
- Gumbel, E. J. (1958). *Statistics of Extremes*. New York, NY: Columbia University Press.
- Hasselmo, M. E., and McClelland, J. L. (1999). *Neural Models of Memory*, Vol. 9, (Boston: Elsevier), 0959–(4388). doi: 10.1016/S0959-438800025-43880027
- Hasson, C., Gaussier, P., and Boucenna, S. (2011). Emotions as a dynamical system: the interplay between the meta-control and communication function of émotions. *J. Behav. Robot.* 2, 111–125.
- Hirel, J., Gaussier, P., Quoy, M., Banquet, J. P., Save, E., and Poucet, B. (2013). The hippocampo-cortical loop: spatio-temporal learning and goal-oriented planning in navigation. *Neural Netw.* 43, 8–21. doi: 10.1016/j.neunet.2013.01.023
- Jauffret, A., Grand, C., Cuperlier, N., Gaussier, P., and Tarroux P. (2013). “How can a robot evaluate its own behaviour? A generic model for self-assessment,” in *International Joint Conference on Neural Networks (IJCNN)* (Dallas, TX).
- Kaski, S., Kangas, J., and Kohonen, T. (1981). Bibliography of self-organising map (SOM) papers: - (1997). *Neural Comput. Surveys* 1, 102–350.
- Kelley, S., Brownell, C., and Campbell, S. (2000). Mastery motivation and self-evaluative affect in toddlers: longitudinal relations with maternal behavior. *Child Dev.* 71, 1061–1071. doi: 10.1111/1467-8624.00209
- Knight, R. T. (1996). Contribution of human hippocampal region to novelty detection. *Nature* 383, 256–259. doi: 10.1038/383256a0
- Kohonen, T., and Oja, E. (1976). Fast adaptive formation of orthogonalizing filters and associative memory in recurrent networks of neuron-like elements. *Biol. Cybern.* 25, 85–95. doi: 10.1007/BF01259390
- Laroque, Ph., Gaussier, N., Cuperlier, N., Quoy, M., and Gaussier, Ph. (2010). Cognitive map plasticity and imitation strategies to improve individual and social behaviors of autonomous agents. *J. Behav. Robot.* 1, 25–36. doi: 10.2478/s13230-010-0004-2
- Lazarus, R. (1991). *Emotion and Adaptation*. New York, NY: Oxford University Press.
- Levine, D. S., and Prueitt, P. S. (1992). “Simulations of conditioned perseveration and novelty preference from frontal lobe damage,” in *Neural Network Models of Conditioning and Action*, Chapter 5, eds M. L. Commons, S. Grossberg, and E. R. J. Staddon (Hillsdale, NJ: Lawrence Erlbaum Associates), 123–147.
- Lewis, D. (2005). Bridging emotion theory and neurobiology through dynamic systems modeling. *Behav. Brain Sci.* 28, 169–194; discussion: 194–245. doi: 10.1017/S0140525X0500004X
- Lisman, J. E., and Otmakova, N. A. (2001). Storage, recall, and novelty detection of sequences by the hippocampus: elaborating on the SOCRATIC model to account for normal and aberrant effects of dopamine. *Hippocampus* 11, 551–568. doi: 10.1002/hipo.1071
- Maillard, M., Gapenne, O., Hafemeister, L., and Gaussier, P. (2005). “Perception as a dynamical sensorimotor attraction basin (2005),” in *Advances in Artificial Life (8th European Conference, ECAL)*. Vol. 3630, (Canterbury), 37–46.
- Marsland, S. (2002). Novelty detection in learning systems. *Neural Comput. Surv.* 3, 1–39.
- Mirolli, M., and Baldassarre, G. (2013). “Functions and mechanisms of intrinsic motivations: The knowledge vs. competence distinction,” in *Intrinsically Motivated Learning in Natural and Artificial Systems*. eds G. Baldassarre and M. Mirolli (Berlin: Springer-Verlag), 49–72.
- O’Keefe, J., and Nadel, N. (1978). *The Hippocampus as a Cognitive Map*. Oxford: Clarendon Press.
- Oudeyer, P.-Y., Kaplan, F., and Hafner, V. (2007). Intrinsic motivation systems for autonomous mental development. *IEEE Trans. Evol. Comput.* 11, 265–286. doi: 10.1109/TEVC.2006.890271
- Pavlov, I. P. (1927). *Conditioned Reflexes: An Investigation of the Physiological Activity of the Cerebral Cortex*. G. V. Anrep, Trans. ed, London: Oxford University Press.
- Prescott, T. J., Gurnett, K., and Redgrave, P. (2001). A computational model of action selection in basal ganglia. i. a new functional anatomy. *Biol. Cybern.* 84, 410.

- Quoy, M., Moga, S., and Gaussier, P. (2003). Dynamical neural networks for top-down robot control. *IEEE Trans. Man Syst. Cybern. A* 33, 523–532. doi: 10.1109/TSMCA.2003.809224
- Richefeu, J., and Manzanera, A. (2006). A new hybrid differential filter for motion detection. *Comput. Vis. Graph.* 32, 727–732.
- Roberts, S., and Tarassenko, L. (1994). A probabilistic resource allocating network for novelty detection. *Neural Comput.* 6, 270–284. doi: 10.1162/neco.1994.6.2.270
- Santucci, V., Baldassarre, G., and Mirolli, M. (2012). “Intrinsic motivation mechanisms for competence acquisition,” in *Development and learning and epigenetic robotics (ICDL), IEEE Int. Conf.* (San Diego, CA), 1–6. doi: 10.1109/DevLrn.2012.6400835
- Scherer, K. R. (1984). “On the nature and function of emotion. A componentprocess approach,” in *Approaches to Emotion*, eds K. R., Scherer, and P. Ekman (Hillsdale: Erlbaum), 293–317.
- Schembri, M., Miroll, M., and Baldassarre, G. (2007). “Evolving childhood’s length and learning parameters in an intrinsically motivated reinforcement learning robot,” in *Proceedings of the Seventh International Conference on Epigenetic Robotics (EpiRob)*. (Lund: Lund University Cognitive Studies), 141–148.
- Schmidhuber, J. (1991). “Curious model-building control system,” in *Proceedings of International Joint Conference on Neural Networks Vol. 2*, (Singapore: IEEE), 1458–1463.
- Schoner, G., Dose, M., and Engels, C. (1995). Dynamics of behavior: theory and applications for autonomous robot architectures. *Robot. Autonom. Syst.* 16, 213–245.
- Sidak, Z., Pranab Sen, K., and Hajek, J. (1967). *Theory of Rank Tests*. 2nd edition (San Diego, CA: Academic Press), 435.
- Stipek, D., Recchia, S., and McClintic, S. (1992). Self-evaluation in young children. *Monogr. Soc. Res. Child. Dev.* 57, 1–98. doi: 10.2307/1166190
- Taylor, S. E., Neter, E., and Wayment H. A. (1995). Self-evaluation processes. *Pers. Soc. Psychol. Bull.* 21, 1278–1287. doi: 10.1177/01461672952112005
- Tronick, Z. (1989). Emotions and emotional communication in infants. *Am. Psychol.* 44, 112–119. doi: 10.1037/0003-066X.44.2.112
- Widrow, B., and Hoff, M. E. Jr. (1960). *Adaptive Switching Circuits*. IWES Convention Record, RE, CON 4, (Stanford, CA), 96–104.
- Zilllich, M., Prankl, J., Morwald, T., and Vincze, M. (2011). “Knowing your limits - self-evaluation and prediction in object recognition, Intelligent Robots and Systems (IROS),” in *IEEE/RSJ International Conference*, (San Francisco, CA), 813–820. doi: 10.1109/IROS.2011.6094856
- Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.
- Received: 21 June 2013; accepted: 15 September 2013; published online: 08 October 2013.*
- Citation: Jauffret A, Cuperlier N, Tarroux P and Gaussier P (2013) From self-assessment to frustration, a small step toward autonomy in robotic navigation. *Front. Neurorobot.* 7:16. doi: 10.3389/fnbot.2013.00016*
- This article was submitted to the journal Frontiers in Neurorobotics.*
- Copyright © 2013 Jauffret, Cuperlier, Tarroux and Gaussier. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.*