information

MDPI

Type of the Paper (Article)

Emergent Robotic Personality Traits via Agent-Based Simulation of Abstract Social Environments

Casey C. Bennett[1,2]

[1] Hanyang University, School of Intelligence Computing, Seoul Korea
[2] DePaul University, College of Computing, Chicago IL
Correspondence: cabennet@hanyang.ac.kr

Abstract: This paper discusses the creation of an agent-based simulation model for interactive robotic faces, built based on data from physical human-robot interaction experiments, to explore hypotheses around how we might create emergent robotic personality traits, rather than pre-scripted ones based on programmatic rules. If an agent/robot can visually attend and behaviorally respond to social cues in its environment, and that environment varies, then idiosyncratic behavior that forms the basis of what we call a "personality" should theoretically be emergent. Here, we evaluate the stability of behavioral learning convergence in such social environments to test this idea. We conduct over 2000 separate simulations of an agent-based model in scaled-down, abstracted forms of the environment, each one representing an "experiment", to see how different parameters interact to affect this process. Our findings suggest that there may be systematic dynamics in the learning patterns of an agent/robot in social environments, as well as significant interaction effects between the environmental setup and agent perceptual model. Furthermore, learning from deltas (Markovian approach) was more effective than only considering the current state space. We discuss the implications for HRI research, the design of interactive robotic faces, and the development of more robust theoretical frameworks of social interaction.

Keywords: Human-Robot Interaction; Robotic Face; Personality, Social Interaction, Agent-Based Simulation; Agent-Based Modeling

Citation: Bennett, C.C.. Emergent Robotic Personality Traits via Agent-Based Simulation of Abstract Social Environments. *Information* **2021**, *12*, x. https://doi.org/10.3390/xxxxx

Academic Editor: Firstname Lastname
Received: date
Accepted: date
Published: date

1. Introduction

1.1. Overview

A primary issue with creating robots and other interactive devices for human interaction is how to create natural seeming behavior from what is otherwise pre-programmed, prescriptive computer code. Naturally intelligent organisms, such as humans, don't all behave the same way. There may be some underlying chemical or instinctive "programming" at play, but the fruition of that is idiosyncratic in terms of individual behavior. Indeed, recent research has shown that even simple sensory systems can produce complex variable reactive behavior in simple organisms [1]. In laymen's terms, we often speak about people or animals as having a *personality*, some stable set of learned and/or innate behavioral patterns unique to the individual. The question is how we might replicate such idiosyncratic behavior in robots and interactive devices in an emergent yet replicable way. These questions have deep implications for human-robot interaction (HRI) as well as the design of robotic faces and virtual avatars.

A number of research efforts have been made toward solving this problem (see Section 1.3 for more details). These largely fall into a couple broad categories. The first approach has been to make use of psychological theories of human personalities, focusing on the "Big 5" personality traits as defined by psychologists [2]. These efforts take as their starting point human-defined labels of these traits – Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism – as a top-down approach based on the assumption that we can create simulated personalities through carefully engineered systems. The second approach has focused on agent-based modeling of personality, using dimension-based or component-based models of personality traits and affect. For instance, cognitive appraisal theory attempts to create cognitive "labels" for different components/dimensions of how environmental stimuli are perceived and responded to in an effort to mimic natural organisms [3]. This is more of a bottom-up approach, though the components/dimensions are still based on human-defined labels of what constitutes a "personality".

The main problem with many of these approaches to personality and affective computing is that the underlying theories they are based on have fundamental conflicts both in a biological sense as well as a computational sense. As Lisetti & Hudlicka (2015) have pointed out [3]:

> *"The different theories often seek to explain different aspects of the overall phenomenon of affect. Consequently, developing an overall theory for affect/emotion modeling would require reconciling not just the theories themselves, narrowly construed, but also their architectural assumptions. This aim, however, resembles the early dreams of strong AI, and its disillusions."*

Their message being that when we take as our starting point human-defined labels of personality, we are relying on the fundamental assumption that it is possible to engineer robust systems of social interaction in AI/robots by examining complex examples of personality in higher organisms and then create computational architectures to emulate them through brute force. This, however, is not presumably how nature evolved the personalities we now see in humans and other higher organisms [1]. Rather, it presumably started with simple variable behaviors across individuals, and somehow molded those over generations into the constructs which we call "personalities" today. An emergent approach. The question then is whether such a process is reproducible in silico. If so, that may reveal important implications toward creating more natural social interaction behavior during human-robot interaction.

Our goal here is to explore these questions via simulation, though one based on an actual physical robotic face. The environment of the simulation is setup to reflect the sensory environment the robotic face experiences while interacting with humans, based on prior physical experiments (see Section 2.1) [4-7].

1.2. Approach

Our approach here takes a different tact. We start with the fundamental definition that the ability to visually attend and behaviorally respond to social cues in its environment lies at the core of any "personality trait" (see Section 2.1). We are not interested in creating human-defined labels for personality, but rather asking the question: can we create an architecture where variable robot behavior results emergently from variations in its social environment, akin to the approach of behavior-based robotics and similar concepts that take of so-called "weak AI" approaches [8,9]? If so, that would suggest that the underlying building blocks of personality might be produced via interactions with the social environment in and of itself, given that that environment is malleable and non-uniform.

In simpler terms: if the social environment varies, and the agent has the capability to attend to and behaviorally respond to social cues, then variable behavior that form the basis of what we call "personality" should be emergent.

The focus here is on exploring *first principles* of how such emergent variable robot behavior (EVRB) might arise during social interaction, rather than engineering a complete robotic platform. However, the simulation experiments described below are based on real-world human interaction experiments with a physical robotic face.

Part of the motivation for this study comes from previous work undertaken a few years prior in an attempt to train neural networks to produce appropriate "life-like seeming" behaviors during interaction between humans and robotic faces, which were not entirely successful, in the sense that the produced models proved to be *unstable*. When the human stimuli were altered from those used previously to learn behavioral patterns, the robot would exhibit erratic behavior, learning and unlearning things in a frenetic manner [6]. Hence our focus here is on the *convergence* toward stable solutions in the face of environmental variation, not just a low training error rate. Learning one specific thing very accurately but which cannot be adapted to the variation seen in real world environments is of limited utility. Indeed, stability is typically seen as one of the key components of "personality" … to the point that unstable personalities are viewed as maladaptive psychological disorders [10,11].

A couple important points should be noted relative to other agent-based modeling approaches to personality and affect (see Section 1.3.2 below). First, averse to more complex approaches to agent emotions based on cognitive appraisal theory, such as the OCC model [12], our approach is pared down to only the valence response to the environment in a pre-cognitive sense, similar to how simple eukaryotes or human infants can respond to stimuli without necessarily knowing what the object stimuli is [1,13]. We do so in order to test how the ability to visually attend and behaviorally respond to social cues might emerge from congruence with environmental stimuli (or inversely the lack thereof) [3]. This ties into our definition for the core of any artificial personality above. Additionally, our approach to agent affect here could also be seen as a pared-down version of Scherer's component process theory (CPT) model, with ours operating only at the sensorimotor level [14]. Similar approaches have been used to control behavior of mobile robots in some previous HRI experiments [15]. We discuss more of this prior work in the next section.

1.3. Background

1.3.1. Robotic Personalities

There have been a number of efforts to develop robotic personalities over the past couple decades. Personality has come to be seen as a core component of human behavior, and thus necessary to emulate in order to create more natural-seeming robotic behaviors [2]. Indeed, even if not explicitly included, people still often attribute personalities to robots, making personality important from a design standpoint for robots in human spaces [16].

Many of these studies have based their development of robotic personalities on the "Big 5" personality traits as defined by psychologists of what "personality" is, for example extraversion and neuroticism (see Section 1.1) [17-23]. Meanwhile other studies have used similar human-defined scales, such as the Myers-Brigg personality scale [2]. A predominant method of evaluation in either case is then to run human-robot interaction experiments and evaluate human perceptions of the robot based on metrics like trust, acceptance, likeability, etc. using some instrument (e.g. the Godspeed scale [24]).

Other research has highlighted some of the problems of simply hard-coding the Big 5 personality traits into robot programming, rather than trying to understand how personality might emerge natively from the underlying architecture. One major limitation is that those studies are in a sense just using robots as a platform to explore theories of human personality traits *mimicked* on a machine, not really studying how robot personalities themselves might be formed [25]. While using robots to study human psychology is certainly a worthwhile endeavor, it is not the only goal we should pursue. Others have criticized the approach on the grounds that hard-coding personality traits does not represent a scalable solution, and that more research needs to be done with an emphasis on scalability [26]. Moreover, truly scalable models of artificial personality will need to be adaptable across culture and context (see Section 1.3.3).

Research on robotic personalities also has a strong link to research on robotic facial expressions, artificial emotions, and other displays of affect in robots and other interactive devices [3]. Many of the studies are based on psychological models of emotion, such as Ekman [27] and Russell [28]. In some HRI research, there is a blurring of the lines between what we might consider mood/affect and personality [29,30]. At the same time, however, psychological research has shown strong interaction effects between mood and personality, and that even in humans it is sometimes difficult to clearly delineate between the two due to the influence they mutually exert on general cognition and behavior [31-33]. We leave it for others to settle that debate. For our purposes here, we take the strong interaction between affect and personality as given, and do not attempt to delineate.

1.3.2. Agent-Based Modeling and Emotion

Along with research on personality and affect in physical robots, there has been a large amount of research done using agent-based modeling. Much of this work has focused more on affect than personality in particular, though regardless it still contains useful insights for designing social interaction and interactive systems [34]. For instance, there has been significant amount of agent-based modeling research into emotional contagion in social groups, and the mechanisms through which affect can spread between individuals much the same as a viral contagion [35,36]. For development of artificial personalities, this underscores the importance that personality and affect play not just in individual behavior, but also in the social communication of information. Personality and affect do not exist in a vacuum.

Other research has focused on how the “desirability” of events in the environment can be used for virtual learning in agents [37]. Some researchers have also used belief-desire-intent (BDI) models to evaluate how *congruence* (or lack thereof) between agent goals and the environment can be used to generate both coherent internal states and agent actions [38,39]. Elsewhere, Gratch and others have begun exploring the deeper implications of agent-based modeling in how we think about the construction of artificial psychologies and development of more natural seeming social interaction between humans and robots, virtual avatars, and other interactive devices [40]. Virtual avatars in particular have been a strong avenue for exploring potential pitfalls when attempting to deploy these approaches during real-world interaction, such as uncanny valley effects that disrupt the virtual experience [41].

There have also been attempts to consolidate all these agent-based simulation approaches for various aspects of cognition and affect, through the creation of cognitive architectures [42]. A prominent example of this is the SOAR cognitive architecture [43]. Research in this vein is still ongoing, but undoubtedly a domain where development of artificial personalities will play a role.

1.3.3. Human-Robot Interaction

From a more general stance of HRI and more broadly human-computer interaction (HCI), the previous sections lead into questions of how we both consider and design interaction, as well as how people relate to devices and technologies in ways that go beyond the physical object itself. As Picard and others have argued, those devices become evocative objects, to which we assign certain anthropomorphic traits that are tied to our identities as human individuals [44]. As much as we shape technology, it shapes us [45].

The term *affective phenomena* includes emotion, affective communication, and personality (as defined by Lisetti & Hudlicka [2]). Simulated models of affective phenomena allow us to test theories of human cognition, and build better interactive systems. Affective systems in artificial agents are thus thought to be critical for creating social fluidity during interaction, as well as for exploring approaches to create more natural goal-conflicted agents that reflect how living organisms have to negotiate their environment in the face of competing demands [3]. *Social fluidity* plays a critical role as well in creating a coherent construct during interaction, which forms the basis of a "virtual experience" [46-48]. Indeed, without a consistent coherent construct, there is no virtual experience, which is why there is sometimes a discrepancy in HRI/HCI between carefully controlled lab studies and less controlled in-the-wild studies [6,46]. There are also cultural and contextual issues at play, which can impact the interaction. This necessitates that our approach to interaction design, including robotic personalities, is adaptable across these variables [49].

Perhaps the most critical advantage of better understanding the formation of artificial personalities in robots and other interactive devices is how it might enhance social cognition of humans during technology interaction, reducing the cognitive load overhead [50-52]. Emotion regulation and its congruence/incongruence with environmental stimuli plays a critical role in human cognition. In fact, a breakdown of that system is thought to underlie some psychological disorders, including personality disorders [53]. In HRI experiments, children interacting with a social robot exhibited a strong innate tendency to create congruence by aligning their behavior temporally with the robot's [54]. As previous research suggests, there are dynamical systems aspects of interaction design – that the environment or interacting agent exhibit some synchrony in response to our actions – which play a core role in both human cognition as well as the overall perceptual experience [5,55,56]. Part of our goal here is to understand the dynamics of the interaction between the agent and its environment, and how those *dynamics* lead to variable behaviors that could underpin the bases of emergent artificial personality [57].

2. Materials and Methods

2.1. Prior Work on Robotic Faces

The simulation experiments described here are based on previous HRI experiments with a physical robot face [4-7]. Those experiments consisted of a camera mounted on an interactive robotic face in order to detect social cues in its environment while interacting with humans. The robot had the ability to convey facial expressions in response to those social cue stimuli, and was mounted on a neck mechanism with 2 degrees-of-freedom (both pan and tilt motion, similar to a human neck) enabling it to track stimuli in its environment (see Table 3 in [4]). Previous research validated the robotic facial expressions, both in-the-lab and in-the-wild experiments in public spaces (Figure 1), across multiple cultural locations: USA, Japan, Germany. Later versions also added vocal speech ability [7]. Here, however, we focus on the visual system of the robot, as it is most pertinent.

(**a**) (**b**)

Figure 1. Examples (a,b) of human-robot interaction during public Art Museum exhibit.

Using video images its onboard camera, the robot calculated dense optical flow to detect motion on 120x120 pixel grid using OpenCV (https://opencv.org/). Points of maximal flow (i.e. motion) were then used as part of a visual attention system (along with face detection done using Haar cascades), with a maximum of four stimuli being held its attention array at any given moment based on evidence of how the human visual attention system works [58]. The robot would then selectively attend to those points in its visual field. An example of this in action can be seen in Figure 2a (note that the participant's face has been deliberately blurred in the image here, for privacy reasons).

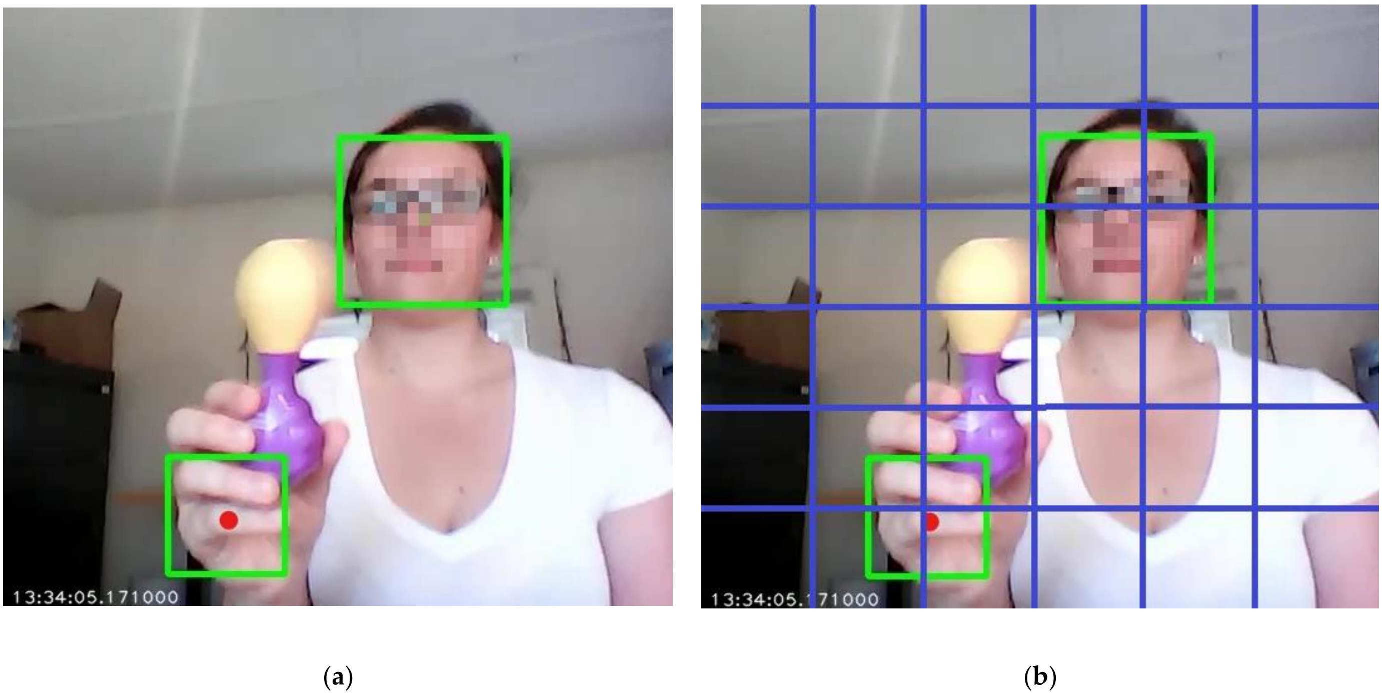


Figure 2. Robotic face perspective during lab-based HRI experiments (a), and 6x6 gridded version representing the sensory environment of the robot used for simulations (b)

The idea here is to create a scaled down version of the attention system of the physical robot in order to create a simplified simulation for testing hypotheses around creating variable robot/agent behavior without explicit programming. We reduce the scale down to a 6x6 grid, which essentially equates to a lower resolution visual system that samples motion at a smaller number of specific points in the image (the center-points of the grid squares). A representation of this can be seen in Figure 2b. Ostensibly, findings from this lower-resolution version should be applicable to a higher-resolution version, just with more computational complexity. Our goal here is to keep things as simple as possible for the simulation, and minimize assumptions needed to be made. More details on the simulation are provided in Sections 2.2 and 2.3 below. The scaled-down system also has the added benefit of mimicking how early visual systems are thought to have evolved from simple collections of a small-number of photo-receptive cells (see Discussion) [1,59].

As mentioned in Section 1.2, our fundamental definition is this: if the social environment varies, and the agent has the capability to attend to and behaviorally respond to social cues, then variable behavior that form the basis of what we call "personality" should be emergent. At a high level, this entails creating an agent that perceives its environment through sensory readings, makes a prediction of how its various potential actions might alter its environment, and how those possible future environments might impact its internal affective state. In short, the agent/robot learns how the world responds to it, not the other way around. The agent has no internal model of the state of the external world, only a model of how its own behaviors are associated with its future affective state. The environment in this sense is simply a *medium* upon which the agent projects its internal state. Learning occurs as the agent observes if the environment responds to it as expected or not. This is a slight, but fundamental, shift in perspective from environment-oriented to agent-oriented that reflects our current scientific understanding of personalities and personality disorders (see Discussion).

By then varying the environmental setup, we should see agents *converge* to variable behavior patterns (or fail to converge in some cases). Such "behavioral stability" is one critical personality trait (see Section 1.2). We detail exactly how this architecture is constructed from a technical standpoint in the next section.

2.2. Simulation Architecture

The simulation here is based on four components: agent, environment, perceptual apparatus, and the learning mechanism. These are described below. In the spirit of scientific replicability, the Python code for these simulation experiments is made available at the authors' website (REDACTED). How the variables in the programming code map to the terms described below is shown in Table 1.

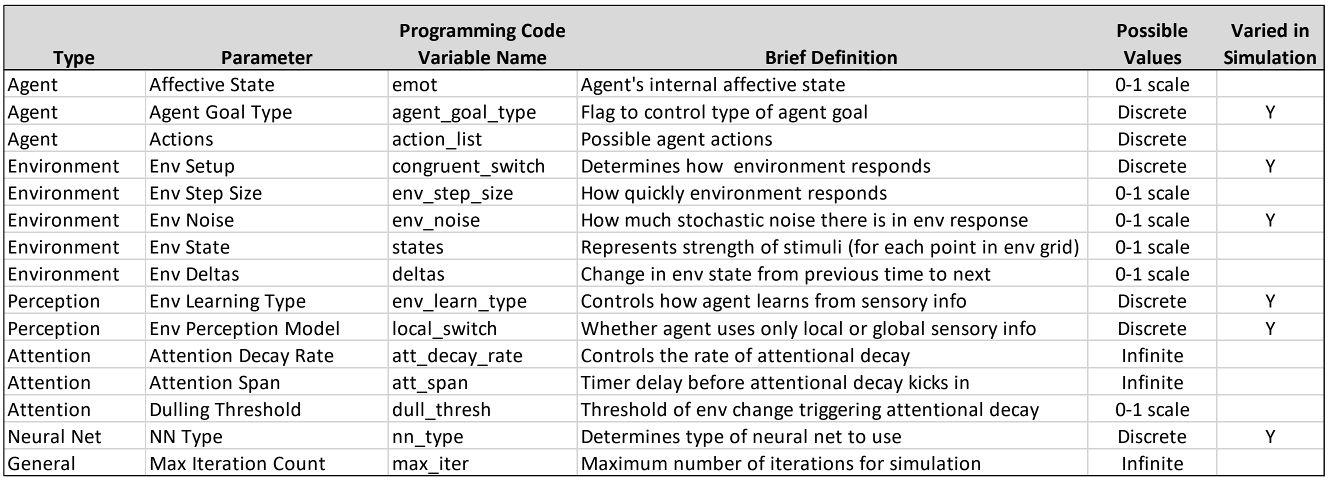
The agent (representing an embodied robotic face) contains a set of internal variables related to its affective state, action choices, attention mechanisms, goal types, and sensory information about the environment. Its internal *affective state* is modeled as a simple 0-1 emotion scale, with 0 representing negative internal emotion and 1 representing positive internal emotion (0.5 in this case can be seen as "neutral"). This is purposely abstract, to model in a simple way that there are some things in our environment we respond negatively to and some we respond positively to (otherwise referred to as emotional valence [27,28]). The action space is also kept abstract, represented as an array of integers from 0-3. The effects of those actions on the environment depend entirely on the environmental setup, described below. The agent also has an attention mechanism which allows it track relevant stimuli in its environment (i.e. social cues). On the physical robot, this was modeled as an ever-changing array of up to 4 stimuli. For simplification purposes in the simulation, we limited it to only one stimuli at a time (though that could be increased in future work), to which we henceforth refer to as the *focal point* of attention. Similar to the physical robot, the simulated agent also had a mechanism for attentional decay. In other words, if a stimuli didn't change or move over time, then the agent's interest would gradually "dull", similar to how attention works in natural organisms [60]. These attention parameters were static here across experiments, and did not change. Finally, the agent needs some sort of overall goal or motivation, otherwise there is nothing to learn as the agent can just behave randomly. We tested three types of goals here relative to the agent's internal affective state: 1) maximizing the time spent in the positive emotion state, 2) maximizing the difference from neutral emotion, whether positive or negative, and 3) having no goal. In plain language, the first goal type could be thought of as pleasure-seeking behavior, and the second more thrill-seeking dramatic behavior. The third goal type (no goal) is a baseline condition where actions are chosen using some basic logic but with no real rhyme or reason. It essentially represents random behavior.

The environment here represents the "sensory environment" of the agent, in this case that reflects the visual field of the physical robotic face (see Section 2.1 for details). This was setup as a 6x6 grid of what the robot "sees". Each point on the grid could take on a value of 0-1, abstractly representing the strength of some sort of stimuli (social cues, motion, etc.). These values would change over time in response to agent actions. These responses can be loosely categorized as congruent, incongruent, or random, based on previous HRI research (see Section 1.3.3) [54]. Each of these variable responses are henceforth referred to as *environmental setups*, some examples can be seen below:

1. Environment congruent response
2. Environment incongruent response
3. Environment random response
4. Environment congruent some percentage of time (25%, 50%, 75%), else random
5. Environment congruent after persistence (same agent action 3 times)
6. Environment congruent above thresholds (e.g. >.7 or <.3)

Here we consider only the first 3 in that list. The congruent response was defined as having the ordinal value of agent actions (e.g. 0,1,2, etc.) align with the direction of change in the environmental response, so that higher action values produced in increase towards 1, lower numbers produced a decrease towards 0, and numbers in the middle producing no change (or a small random fluctuation around the current value). The incongruent response was the opposite. Random response was, of course, completely random. The size of the response was scaled by the distance from the current focal point, so that as the distance from the current focal point increased the effect inversely decreased (based on Euclidean distance). The idea here being that in a social environment, the effects of your actions should have their primary effect on whatever social stimuli you are attempting to interact with. Other environmental parameters included an environmental *step size*, which represented how quickly the environment changed in response to agent actions, along with an *environmental noise* factor, which controlled whether the environmental response was stochastic or not (and to what degree). Both of those parameters could take on values on a scale of 0-1, but where not specified below were kept at small values (0.1 and 0.05, respectively).

In order to interact with the environment of course, the agent also needs some perceptual apparatus. The agent maintained internal sensory readings about its environment (one data point from each point of the grid). This was modeled in two ways: 1) sensory readings of the current state, and 2) deltas (i.e. difference) of the current state from the previous state. These represent a static vs. Markovian approach, respectively [61]. We henceforth refer to these two models of sensory perception as *environmental learning types*. We assume for simplicity that the agent can always see its entire visual field at any given time. However, the agent's perception could also vary depending on whether it only considered local changes near its current focal point (which might be more relevant since they are occurring close to the stimuli), or always considered global changes from its entire environment (i.e. visual field in this case). Here, local was defined as the focal point itself and its immediate neighbors on the grid (see Figure 2), with global being all points on the grid. Hence we tested different *environmental perception models* (local vs. global).



Type	Parameter	Programming Code Variable Name	Brief Definition	Possible Values	Varied in Simulation
Agent	Affective State	emot	Agent's internal affective state	0-1 scale	
Agent	Agent Goal Type	agent_goal_type	Flag to control type of agent goal	Discrete	Y
Agent	Actions	action_list	Possible agent actions	Discrete	
Environment	Env Setup	congruent_switch	Determines how environment responds	Discrete	Y
Environment	Env Step Size	env_step_size	How quickly environment responds	0-1 scale	
Environment	Env Noise	env_noise	How much stochastic noise there is in env response	0-1 scale	Y
Environment	Env State	states	Represents strength of stimuli (for each point in env grid)	0-1 scale	
Environment	Env Deltas	deltas	Change in env state from previous time to next	0-1 scale	
Perception	Env Learning Type	env_learn_type	Controls how agent learns from sensory info	Discrete	Y
Perception	Env Perception Model	local_switch	Whether agent uses only local or global sensory info	Discrete	Y
Attention	Attention Decay Rate	att_decay_rate	Controls the rate of attentional decay	Infinite	
Attention	Attention Span	att_span	Timer delay before attentional decay kicks in	Infinite	
Attention	Dulling Threshold	dull_thresh	Threshold of env change triggering attentional decay	0-1 scale	
Neural Net	NN Type	nn_type	Determines type of neural net to use	Discrete	Y
General	Max Iteration Count	max_iter	Maximum number of iterations for simulation	Infinite	

Table 1: Parameter Mappings

Changing the environment, and how the agent perceived the environment, forms the basis of our experiments below. The goal is to show that the behavior of the agent will converge to different behavioral patterns (in terms of learning stability), without being explicitly programmed to do so.

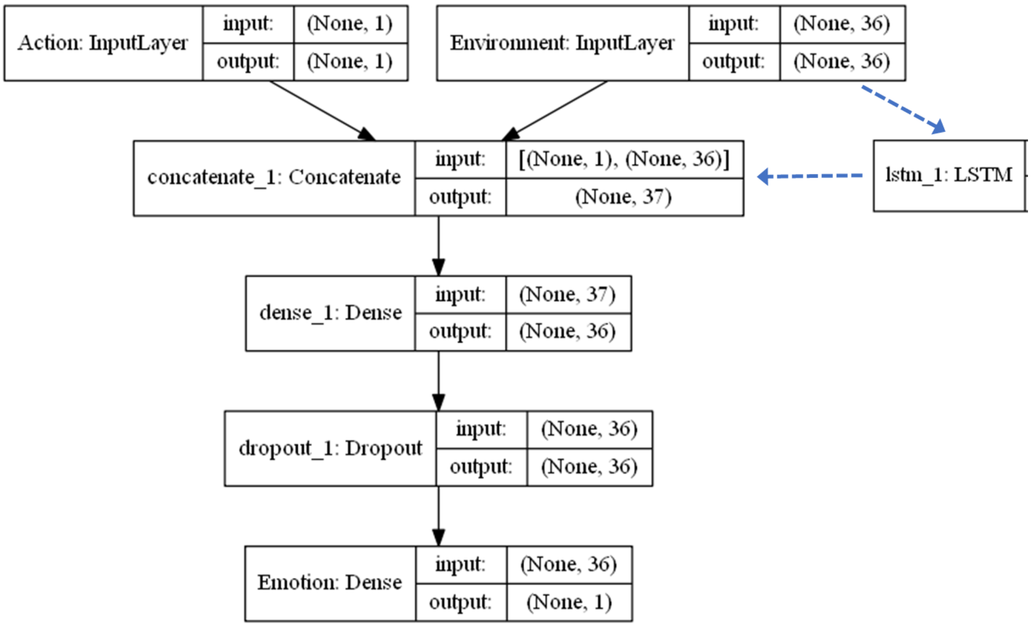
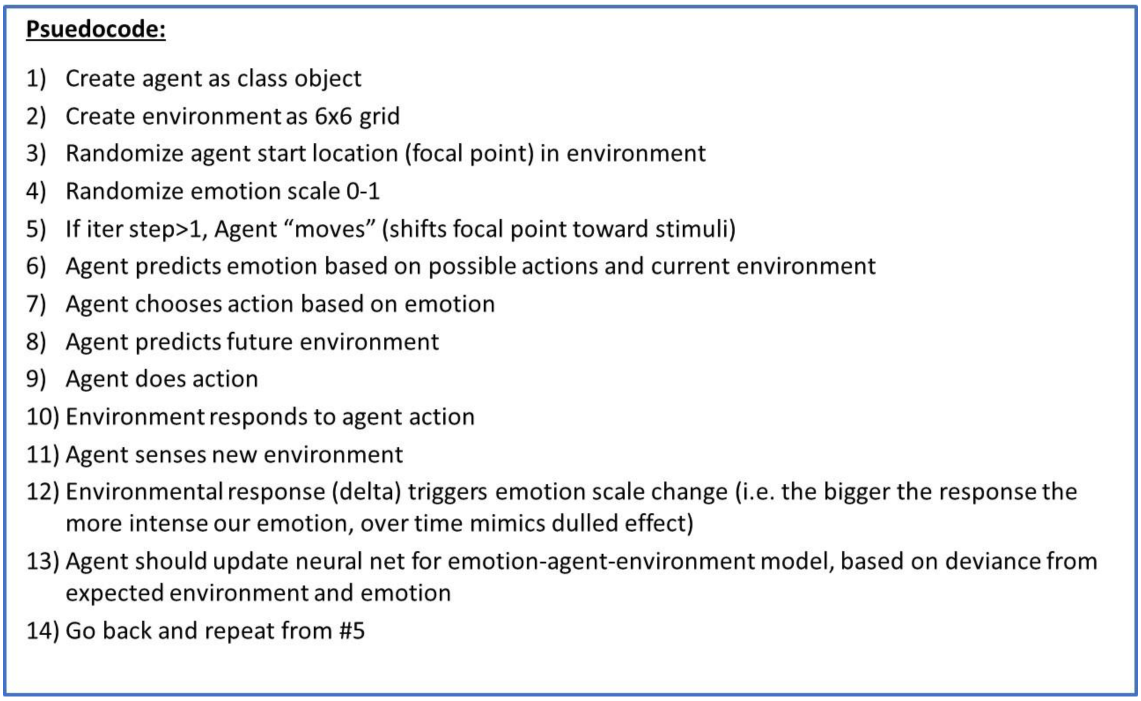


Figure 3. MLP Neural Network Schematic. Optional LSTM module for RNN version showed in blue dotted lines, with the input shape taking form (None, 2, 36) for the simple case of current and previous timestep.

The learning mechanism here was operationalized as a neural network, using the Keras package in python running on top of TensorFlow (https://keras.io/). We experimented with a couple different neural network architectures: 1) a simple feed-forward multi-layer perceptron (MLP), and 2) a recurrent neural network (RNN) with recurrent connections to give the agent a sort of "memory" of what it had previously seen (via a long-short term memory, or LSTM [62]). To be completely clear, this paper is not intended as an evaluation or comparison of different neural network models. As such, the models were kept relatively plain vanilla, with an input layer with nodes for each environmental sensory reading and possible agent actions (n=37), a single hidden layer that was half the size of the input layer, and an output layer to with a single node to predict the agent's future affective state. A visual representation of the MLP network is shown in Figure 3. The RNN version is the same, just with an additional LSTM layer acting on the environment input prior to the concatenate step. Three functions were written for each neural network type: 1) a construct function to initialize the network, 2) an update function to feed information in for learning at each iteration of the simulation run, 3) and a predict function so the agent could call on the neural network for predictions in order to choose an action. All the code for this is included simulation Python code available on the authors' website, as mentioned in the first paragraph of this section. There are parameters in the code to explore the effects of things like increasing the number of hidden layers and incorporating dropout layers, but those are not part of the experiments here. We discuss the potential to utilize more complex neural network models in future work in the Discussion section.

2.3. Experimental Design

The simulation itself can be represented by the pseudo-code shown in Figure 4. In short, after initializing the agent and environment, the agent goes through a series of iterations. On each iteration, the agent "moves" (which here would represent a shift in its gaze direction) based on detected stimuli. It then predicts what will happen in the future given various actions it could take (using its internal neural network), both how the environment will change and the impact of that on its internal affective state. After that, the agent chooses an action to perform, and the environment responds. Finally, the agent senses the new environment, which triggers internal affective state changes, and uses the deviance from the predicted environment and affective state to update its model of how the world responds to it (represented as a neural network).



Psuedocode:

```
1)  Create agent as class object
2)  Create environment as 6x6 grid
3)  Randomize agent start location (focal point) in environment
4)  Randomize emotion scale 0-1
5)  If iter step>1, Agent "moves" (shifts focal point toward stimuli)
6)  Agent predicts emotion based on possible actions and current environment
7)  Agent chooses action based on emotion
8)  Agent predicts future environment
9)  Agent does action
10) Environment responds to agent action
11) Agent senses new environment
12) Environmental response (delta) triggers emotion scale change (i.e. the bigger the response the
    more intense our emotion, over time mimics dulled effect)
13) Agent should update neural net for emotion-agent-environment model, based on deviance from
    expected environment and emotion
14) Go back and repeat from #5
```

Figure 4. Simulation Pseudocode.

A few additional details. For our main series of experiments, the maximum iteration count was set to 1000, though for some subsequent experiments that was increased to 10000. For the affective state, it was updated based on how the environment changed from the previous step to the next (i.e. the environmental "delta"), which varied depending on the environmental setup (see Section 2.2). The information used to calculate that environmental delta was determined by the environmental perceptual model (local vs. global). That delta value was then added to the affective state, either increasing it (towards 1) or decreasing it (towards 0). If the agent's attention span had been exceeded, the affective state would decay back towards neutral (0.5).

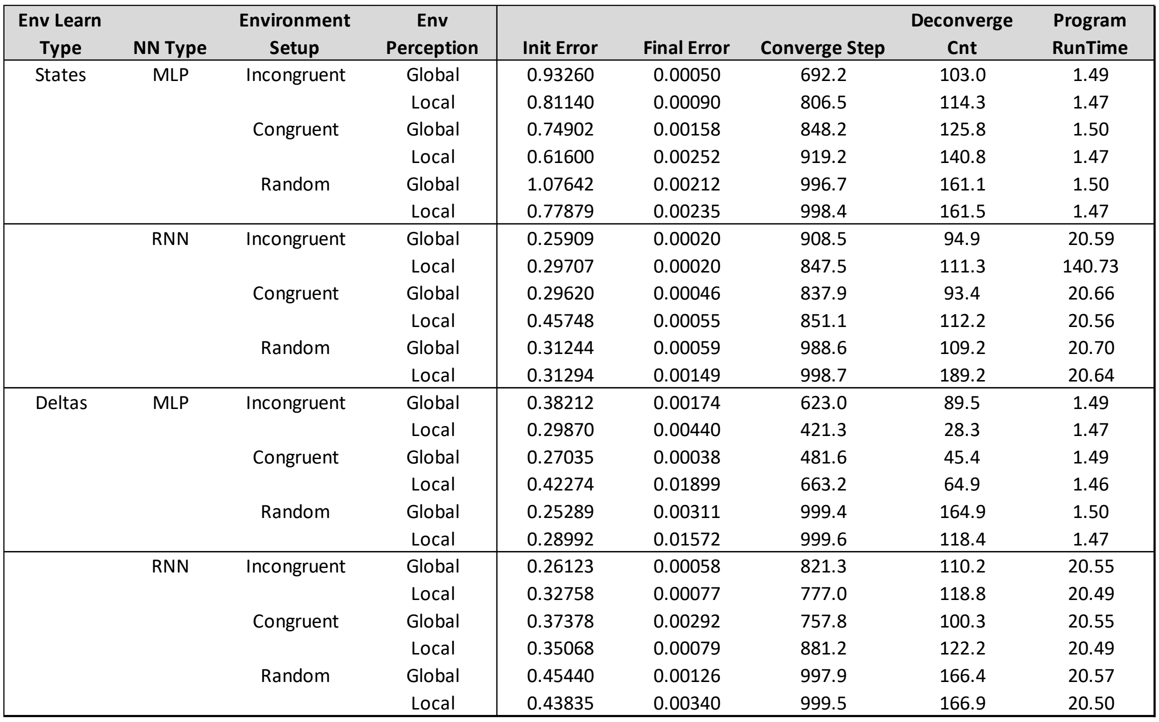
Five of the parameters defined in Table 1 were varied in our experiments: Environmental Setup, Environment Learning Type, Environment Perception Model, Agent Goal Type, and Neural Network Type. Different combinations of each parameter setting formed a "permutation". Each permutation was repeated 30 times in order to estimate average performance metrics, for a total of 2,160 individual simulation runs for the main results. The simulation analysis was exploratory.

Our evaluation metric here for each experimental hypotheses is taken as the "learning convergence". We define this two-fold: 1) how quickly the robot learns how the environment responds to its behaviors, and 2) how stable that solution is. These are henceforth referred to as *converge step* and *deconverge count*, respectively. Converge step represents the iteration step where the agent reached final learning convergence in that experiment. Deconverge count represents the number of times the agent's learned model deconverged prior to final convergence. Here, convergence was defined as when the change in mean squared error (MSE) from the previous iteration to the next dropped below 0.001. Experimental permutations that converged more quickly and with less instability (deconvergence) were taken as more successful, whereas permutations that never converged without subsequent deconvergence were assessed as unsuccessful. Beyond the main experiments, we conducted additional experiments on the effects of environmental noise, as well longer max iteration times.

3. Results

3.1. Main Results

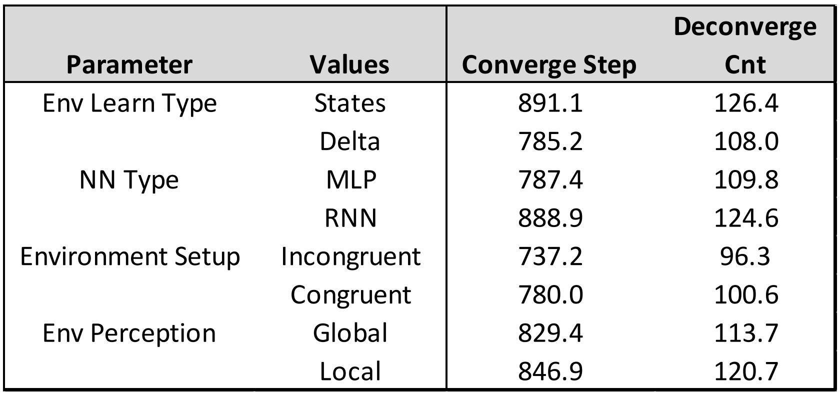
For reference, the full results of the main experiments are included in Table A1 in the appendix. The results are broken out by agent goal type. We note that as expected, agents with "no goal" failed to achieve learning convergence (based on converge step and deconverge cnt), while at the same time agents with the "different from neutral" goal also failed to stably converge. For reasons outlined in previous sections, we take as stable convergence as our primary metric here. As such we focus in this paper on the results using the "max positive" goal. The main results for that agent goal type are shown in Table 2.



Env Learn Type	NN Type	Environment Setup	Env Perception	Init Error	Final Error	Converge Step	Deconverge Cnt	Program RunTime
States	MLP	Incongruent	Global	0.93260	0.00050	692.2	103.0	1.49
			Local	0.81140	0.00090	806.5	114.3	1.47
		Congruent	Global	0.74902	0.00158	848.2	125.8	1.50
			Local	0.61600	0.00252	919.2	140.8	1.47
		Random	Global	1.07642	0.00212	996.7	161.1	1.50
			Local	0.77879	0.00235	998.4	161.5	1.47
	RNN	Incongruent	Global	0.25909	0.00020	908.5	94.9	20.59
			Local	0.29707	0.00020	847.5	111.3	140.73
		Congruent	Global	0.29620	0.00046	837.9	93.4	20.66
			Local	0.45748	0.00055	851.1	112.2	20.56
		Random	Global	0.31244	0.00059	988.6	109.2	20.70
			Local	0.31294	0.00149	998.7	189.2	20.64
Deltas	MLP	Incongruent	Global	0.38212	0.00174	623.0	89.5	1.49
			Local	0.29870	0.00440	421.3	28.3	1.47
		Congruent	Global	0.27035	0.00038	481.6	45.4	1.49
			Local	0.42274	0.01899	663.2	64.9	1.46
		Random	Global	0.25289	0.00311	999.4	164.9	1.50
			Local	0.28992	0.01572	999.6	118.4	1.47
	RNN	Incongruent	Global	0.26123	0.00058	821.3	110.2	20.55
			Local	0.32758	0.00077	777.0	118.8	20.49
		Congruent	Global	0.37378	0.00292	757.8	100.3	20.55
			Local	0.35068	0.00079	881.2	122.2	20.49
		Random	Global	0.45440	0.00126	997.9	166.4	20.57
			Local	0.43835	0.00340	999.5	166.9	20.50

Table 2: Main Results (Max Pos Goal Only)

There are some notable differences in the results, in particular a few rows in the middle of Table 2 (in the Deltas+MLP section) that significantly outperformed most other models. We highlight a few general trends in these results. First, the environmental setup did impact the agent's ability to learn. An environment that responds randomly to agent actions, not surprisingly, failed to produce stable convergence, while both incongruent and congruent environmental responses could be learned. Second, the optimal information for the agent to utilize from the environment (i.e. environmental perceptual model) depended on the environmental setup. For congruent response environments, it was better to utilize global information. For incongruent response environments, it was better to use more local information near the focal point. In layman's terms, one could perhaps think of this as a greater necessity of minimizing distractions in incongruent environments. We also note that learning using deltas (how the environment was changing), rather than absolute values of the environment's current state, was more effective in achieving stable learning convergence. Finally, although this paper is not focused on exploring different neural network architectures, we do note that MLPs outperformed RNNs, though that may not hold true in larger, more complex sensory environments than that used in these experiments.



Parameter	Values	Converge Step	Deconverge Cnt
Env Learn Type	States	891.1	126.4
	Delta	785.2	108.0
NN Type	MLP	787.4	109.8
	RNN	888.9	124.6
Environment Setup	Incongruent	737.2	96.3
	Congruent	780.0	100.6
Env Perception	Global	829.4	113.7
	Local	846.9	120.7

Table 3: Parameter Comparison

A head to head comparison of some of the parameters in *isolation* can be seen in Table 3. We do note that this glosses over some of the interaction effects that can be seen in Table 2. We tested for significant differences in each of the main parameter settings using a fixed-effects 4-way ANOVA in R statistical software for both evaluation metrics (converge step and deconverge count). Results are shown in Tables 4 and 5 (note parameter names here are using the "programming code variable names" shown in Table 1). All the parameters exerted main effects on one or both of the evaluation metrics, though the environmental perception model was the weakest in that sense. However, both the environmental setup (congruent.switch) and perceptual model (local.switch) also exerted significant impact through interaction with *each other* and other parameters. We also detected a significant interaction between all four parameters for the deconverge count. Additionally, we stripped out the NN type parameter looking only at the MLP neural network results, then reran the ANOVAs, and found the same pattern of effects (data not shown for brevity). The main takeaway from Tables 3-5 in comparison to Table 2 is just how important the interaction effects are here between the parameters in order to explain the detailed minutiae of patterns seen in Table 2.

Parameter	Df	Sum Sq	Mean Sq	F value	Pr(>F)	Sign. Level
Main Effects						
Env.Learn.Type	1	2018136	2018136	23.75	<.001	***
NN.Type	1	1853289	1853289	21.81	<.001	***
Congruent.Switch	1	8122924	8122924	95.58	<.001	***
Local.Switch	1	55178	55178	0.65	0.421	
Interaction Effects						
Env.Learn.Type:NN.Type	1	957834	957834	11.27	<.001	***
Env.Learn.Type:Congruent.Switch	1	735081	735081	8.65	0.003	**
NN.Type:Congruent.Switch	1	1262596	1262596	14.86	<.001	***
Env.Learn.Type:Local.Switch	1	9776	9776	0.12	0.735	
NN.Type:Local.Switch	1	19313	19313	0.23	0.634	
Congruent.Switch:Local.Switch	1	79877	79877	0.94	0.333	
Env.Learn.Type:NN.Type:Congruent.Switch	1	158268	158268	1.86	0.173	
Env.Learn.Type:NN.Type:Local.Switch	1	132167	132167	1.56	0.213	
Env.Learn.Type:Congruent.Switch:Local.Switch	1	156891	156891	1.85	0.175	
NN.Type:Congruent.Switch:Local.Switch	1	1442	1442	0.02	0.896	
Env.Learn.Type:NN.Type:Congruent.Switch:Local.Switch	1	216495	216495	2.55	0.111	
Residuals	704	59831846	84988			

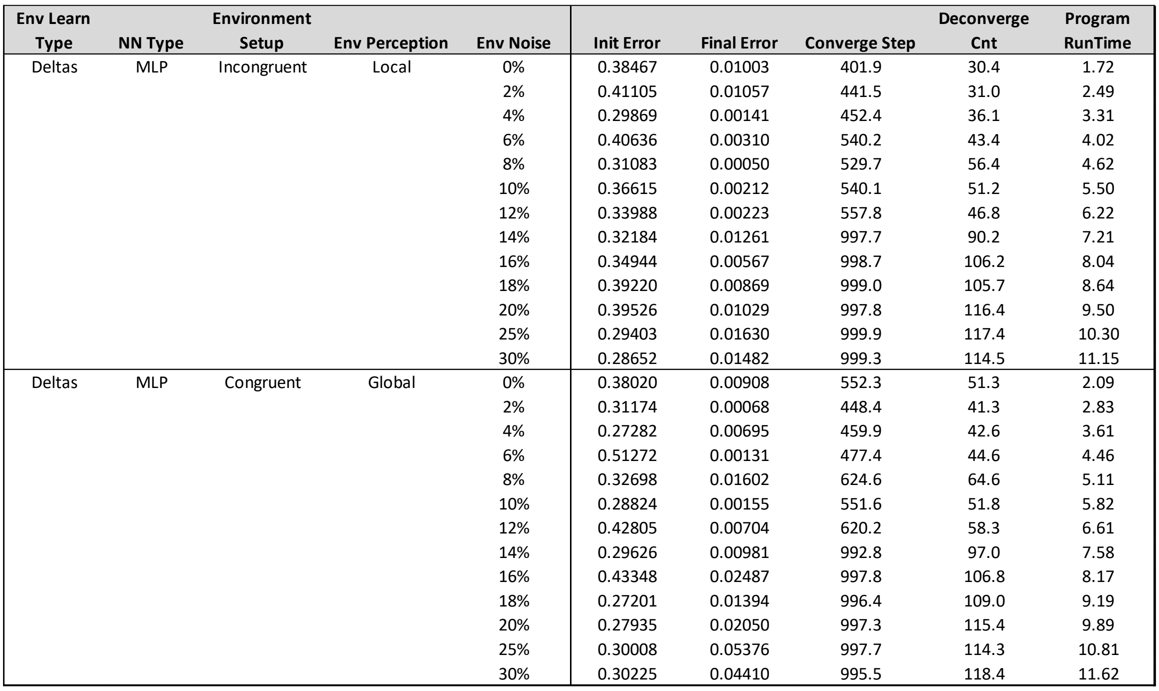
Table 4: Parameter effects ANOVA – Converge Step

Parameter	Df	Sum Sq	Mean Sq	F value	Pr(>F)	Sign. Level
Main Effects						
Env.Learn.Type	1	60830	60830	27.78	<.001	***
NN.Type	1	39043	39043	17.83	<.001	***
Congruent.Switch	1	409326	409326	186.93	<.001	***
Local.Switch	1	8946	8946	4.09	0.044	*
Interaction Effects						
Env.Learn.Type:NN.Type	1	170755	170755	77.98	<.001	***
Env.Learn.Type:Congruent.Switch	1	9819	9819	4.48	0.035	*
NN.Type:Congruent.Switch	1	10370	10370	4.74	0.030	*
Env.Learn.Type:Local.Switch	1	49601	49601	22.65	<.001	***
NN.Type:Local.Switch	1	53941	53941	24.63	<.001	***
Congruent.Switch:Local.Switch	1	6593	6593	3.01	0.083	.
Env.Learn.Type:NN.Type:Congruent.Switch	1	4326	4326	1.98	0.160	
Env.Learn.Type:NN.Type:Local.Switch	1	1181	1181	0.54	0.463	
Env.Learn.Type:Congruent.Switch:Local.Switch	1	4008	4008	1.83	0.177	
NN.Type:Congruent.Switch:Local.Switch	1	5038	5038	2.30	0.130	
Env.Learn.Type:NN.Type:Congruent.Switch:Local.Switch	1	17751	17751	8.11	0.005	**
Residuals	704	1541569	2190			

Table 5: Parameter effects ANOVA – Deconverge Cnt

3.2. Effects of Environmental Noise

Beyond the main results, another question was how much stochastic noise could be present in the environmental response before agent learning broke down. Ostensibly, if an agent/robot cannot reliably detect environmental responses due to noise – whether that be stochastic processes in the environment or measurement noise in its own sensory detectors – then learning should be disrupted. To explore this, the environmental noise parameter was varied between 0 and 30% (0.3). The results can be seen in Table 6.



Env Learn Type	NN Type	Environment Setup	Env Perception	Env Noise	Init Error	Final Error	Converge Step	Deconverge Cnt	Program RunTime
Deltas	MLP	Incongruent	Local	0%	0.38467	0.01003	401.9	30.4	1.72
				2%	0.41105	0.01057	441.5	31.0	2.49
				4%	0.29869	0.00141	452.4	36.1	3.31
				6%	0.40636	0.00310	540.2	43.4	4.02
				8%	0.31083	0.00050	529.7	56.4	4.62
				10%	0.36615	0.00212	540.1	51.2	5.50
				12%	0.33988	0.00223	557.8	46.8	6.22
				14%	0.32184	0.01261	997.7	90.2	7.21
				16%	0.34944	0.00567	998.7	106.2	8.04
				18%	0.39220	0.00869	999.0	105.7	8.64
				20%	0.39526	0.01029	997.8	116.4	9.50
				25%	0.29403	0.01630	999.9	117.4	10.30
				30%	0.28652	0.01482	999.3	114.5	11.15
Deltas	MLP	Congruent	Global	0%	0.38020	0.00908	552.3	51.3	2.09
				2%	0.31174	0.00068	448.4	41.3	2.83
				4%	0.27282	0.00695	459.9	42.6	3.61
				6%	0.51272	0.00131	477.4	44.6	4.46
				8%	0.32698	0.01602	624.6	64.6	5.11
				10%	0.28824	0.00155	551.6	51.8	5.82
				12%	0.42805	0.00704	620.2	58.3	6.61
				14%	0.29626	0.00981	992.8	97.0	7.58
				16%	0.43348	0.02487	997.8	106.8	8.17
				18%	0.27201	0.01394	996.4	109.0	9.19
				20%	0.27935	0.02050	997.3	115.4	9.89
				25%	0.30008	0.05376	997.7	114.3	10.81
				30%	0.30225	0.04410	995.5	118.4	11.62

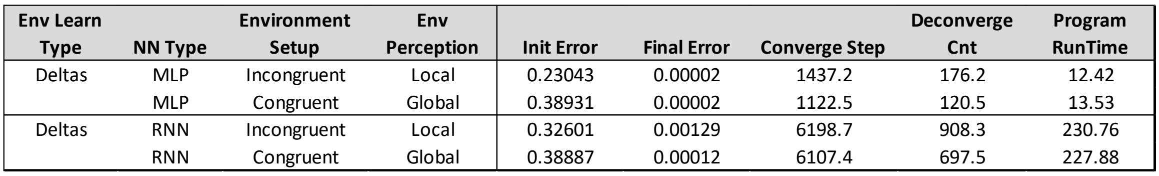
Table 6: Environmental Noise Effects

The effects appear to be fairly stable, with some oscillation down below 10%. Convergence values do gradually rise, and there appears to be a sharp demarcation point around 12-14% beyond which environmental noise causes the learning process to completely breakdown. The specific numbers are interesting here, but they may merely be paradigmatic of the parameters and settings used in this particular simulation. A different construction or more complex sensory environment may produce different specific numbers. More research is needed in the future to explore that question. However, for our purposes here, we did find that environmental noise in a simulation of social sensory environment does have the potential to impact/disrupt learning convergence in an agent/robot. Accounting for environmental noise appears to be an important component for modeling agent learning in a social environment.

3.3. Extending the Maximum Iteration Limit

For a few of the higher performing models from the main results above, we were curious what if any effect might occur over longer iteration times. How often would the agent de-converge at a later point? Would the longer iteration times allow for recurrence in the neural networks to have more of an effect (leading to better RNN performance)?

To explore these hypotheses, we increased the max iteration limit to 10000 for two of the best performing environments in Table 2 above: the congruent global version and the incongruent local version (both using deltas as for environmental learning). This was done for both the MLP and RNN networks, resulting in 4 total experiments. Results can be seen in Table 5.



Env Learn Type	NN Type	Environment Setup	Env Perception	Init Error	Final Error	Converge Step	Deconverge Cnt	Program RunTime
Deltas	MLP	Incongruent	Local	0.23043	0.00002	1437.2	176.2	12.42
	MLP	Congruent	Global	0.38931	0.00002	1122.5	120.5	13.53
Deltas	RNN	Incongruent	Local	0.32601	0.00129	6198.7	908.3	230.76
	RNN	Congruent	Global	0.38887	0.00012	6107.4	697.5	227.88

Table 7: Extending the Max Iteration Limit

As we can see, there is now a greater spread in terms of converge step and deconverge count, although the general relative pattern of results was the same as seen in Table 2. The main takeaway from these experiments was that increasing the max iteration limit did not substantively change the patterns of learning stability. Nor did it lead to better RNN performance. As mentioned in Section 3.1, this may not, however, hold true in more complex sensory environments.

3.4. Algorithmic Analysis

Another critical aspect of any model attempting to manifest artificial personalities and/or personality is understanding the computational complexity required for encoding the information into computer system, robot, virtual avatar, etc. Such an analysis will be important for comparing different models, and understanding resource needs. However, computational complexity for a model such as the one used in this paper will heavily depend on the choice of neural network architecture, which can vary widely. A more generalizable approach is to evaluate such a system based on principles of information theory and the branching factor of the search space [63].

As such, we propose to use decision-theoretic approach (based on Shannon's Information Theory [64] to analyze this, which describes the number of bits needed to encode the model based on the number of parameters and possible parameter settings. Given Shannon's equation:

$$H(X) = \sum_{i=1}^{n} p(x_i)I(x_i) = -\sum_{i=1}^{n} p(x_i)\log_b p(x_i) \quad (1)$$

Where the joint entropy $p(x_i)$ is 1/n all the parameter i, where n is the number of possible settings for that parameter, and b is set to 2 to calculate binary bits. Totaling $H(x)$ for the various probabilities for our simple model here equates to 6.755 bits. Such an approach could be used to compare models of artificial personality in the future based on their information content (encoded bits) and measurable performance. Ideally, we would prefer both high performing models and less complex, parsimonious models (lower information content), but likely there will be some tradeoff between performance and complexity.

4. Discussion

4.1. Summary of Key Findings

This paper describes the creation of an agent-based simulation model for interactive robotic faces, built based on data from real-world physical human-robot interaction experiments. We use this abstract simulation to explore hypotheses around how we might create the building blocks of emergent robotic personality traits, rather than pre-scripted ones based on programmatic rules. We ran a number of experiments based on those hypotheses, evaluating how variations in the social environment of the robot can produce variations in learning behavior in the simulated robot. The main findings can be summarized as follows:

1. Altering the environment, while holding the robot agent constant, can alter the robot's behavior and learning patterns (Table 2). In particular, certain environments were notably better in producing stable behavioral learning convergence, whereas others produced only chaotic behavior. This appears to support our primary hypothesis from Section 1.2.
2. The environment here varied based on how it responded to the agent (environmental setup), as well as how it was perceived by the agent (environmental perceptual model). There were significant interaction effects between the two as well (Tables 3 and 4).
3. For Congruent response environments, it is better to utilize global information. For Incongruent response environments, it is better to use more local information near the focal point of the visual field. This can be seen in the interaction patterns between the *environmental perception* and *environmental setup* parameters in Table 2. In layman's terms, one could perhaps think of this as a greater necessity of minimizing distractions in incongruent environments.
4. The robot agent was much more effective at learning how to respond using environmental deltas (difference of the current state from the previous state), rather than the absolute environmental states themselves (current information only), as shown in Table 2. In other words, a Markovian approach seemed more effective.
5. These effects are dependent on the robot agent having some sort of "goal", even just a simple goal like maximizing their positive emotion level (Table A1).
6. In terms of neural networks, MLPs in general performed better than RNNs. This should be taken with caution, however. That may or may not be true, however, in more complex sensory environments.
7. Environmental noise has a direct effect on this. Too much noise in the environmental response (e.g. measurement noise from its own sensors) disrupts the agent's ability to achieve stable learning convergence (Table 6). Accounting for environmental noise appears to be an important component for modeling agent learning in a social environment.
8. Extending the maximum iteration steps to a higher limit did not appear to alter the above patterns (Table 7).

4.2. HRI Implications

The main findings above suggest a number of interesting broader implications for HRI, beyond the results themselves. The first relates back to our arguments in Section 1.2, around how the underlying building blocks of personality might be produced via interactions with the social environment in and of itself. Can we create variable robotic personalities without explicitly programming them? The results here seem to suggest – given an agent/robot that learns how the environment responds to it and a variable environment – that there may be systematic dynamics in the learning patterns of the agent. Those dynamics lead to variable behaviors, which could underpin the bases of emergent artificial personality [5,55,56]. This holds potential for HRI researchers to take advantage of when designing interaction. In other words, what we may need is not more clever programming, but a better understanding of the dynamics, in order to produce a scalable approach to idiosyncratic robotic social behavior and robotic personalities.

Second, there were significant performance improvements in learning behavior when using delta information (differences from previous state) rather than absolute state values. This suggests a potential connection to Markov models in terms of temporal learning for artificial personalities. Indeed, previous AI studies have shown the importance of deltas for modeling temporal patterns [61]. When attempting to learn how to behave temporally, there are limits to what can be gleaned from the current sensory environment, and what may be more critical is for the agent/robot to understand how the world is changing, rather than attempting to build a model of the current state of the world. There is evidence for that sort of process (termed dynamic predictive coding) in the functioning of biological human brains [65,66], as well as in previous HRI studies [67]. The relative character of Markovian information also holds advantages over static information in that it more naturally produces invariant representations, which can recognize situations that are essentially the same despite slight permutations in appearance [68,69]. This finding may hold significance for future HRI experiments in terms of designing robot perception in social environments.

Third, the scaled-down approach taken here holds potential for helping us understand the mechanics of social interaction from a more fundamental level, and how sensory systems and social behavior co-evolved. Evolution didn't start with fully formed social mechanisms or "personalities". Rather, it presumably started with simple variable behaviors across individuals, and somehow molded those over generations into the constructs which we call "personalities" today. The scaled-down system mimics how early visual systems are thought to have evolved from collections of a small-number of photo-receptive cells [59]. Recent research has shown that even such simple sensory systems can produce complex variable reactive behavior in simple organisms [1]. Similar simplified systems have also been developed to manifest approach/avoidance in robots using a small set of parameters [70,71]. For instance, Jones et al. (2014) showed how complex robot interaction behavior could arise from a simple control model based on epigenetic hormone switches that up or down regulate in response to the environment [71]. In that sense, developing a more fundamental understanding of the building blocks of complex social behaviors in higher organisms such as humans and how they developed will allow HRI researchers to create more robust theoretical frameworks of social interaction, which will enable us to design more precise experiments targeting specific phenomena (e.g. personality components, affective state, cultural aspects) [3,49].

Finally, much psychological investigation into human personalities focuses on pathological cases, otherwise known as personality disorders. Studying pathological cases – places where normal personality manifestations break down – can provide useful clues to where "personality" and/or a "sense of self" comes from in natural cognitive systems. Indeed, those break downs, and their developmental etiology, point to critical aspects of personality and the formation thereof [10,72,73]. On the flip side, attempting to engineer interactive robots and other agents based on such psychological research may also provide a better understanding of how human personalities work, generating new testable hypotheses for future psychological experiments. The HRI research presented here is one example of that.

4.3. Limitations

There are a number of limitations to the research presented here. First of all, the abstract aspects of the model (actions, emotions, environment, etc.) means that we should be cautious about over-interpreting the results. There are an endless array of different parameter settings that could be explored, and here we only explored a subset of them due to practicality. Along with that, when the simulated models are ported back to the physical robot in future work (see next section), we may find differences between the simulation and reality. The same issues apply to neural network architectures used here. For pragmatic purposes, we simply chose an architecture and used it. Evaluating/comparing neural network performance was not our goal here. That said, future research may find differences depending on the architecture used. That remains to be seen, and is beyond the scope of this work. Finally, one major limitation (described in Section 1.3.1) is that is no clear delineation between some components of social interaction (e.g. personality and affective state) currently exists in the literature. As such, we did not attempt to make that delineation here. However, more robust theoretical frameworks of social interaction in HRI are needed in the future to design more precise experiments.

We also would re-emphasize that this paper is not intended as an evaluation or comparison of different neural network models, so any strong conclusions in that regard should not be drawn. The models here were kept simplistic by design, and more complex sensory environments or more sophisticated architectures might reveal important differences between different types of neural networks (MLPs, RNNs, LSTMs, etc.), particularly in terms of convergence dynamics. We leave that as an open question for future work.

4.4 Future Work

There are a several avenues of potential future work related to the study presented here. We present a few of those below:

1. **Implementing the Simulation Model on the Physical Robotic Face**: Since the agent-based simulation is based on data from actual robotic face HRI interaction experiments, it could be ported back to the physical robot for future HRI studies to compare the simulation with physical reality.
2. **Exploring Other Neural Network Architectures**: In particular, as the main sensory mode here is visual, it could be interesting to explore the use of convolutional neural networks, especially with the full-scale visual images. In the same sense, adding autoencoders may be useful to compress information in more complex sensory environments.
3. **Exploring Other Environmental Setups**: This might entail adding additional variables to define the environment, or exploring larger state spaces or action spaces. More complex variations of congruence/incongruence could be implemented. Additionally, other types of agent goals could be modeled.
4. **Learned Attention Parameters:** The parameters of the attentional mechanism could be learned from interaction with the environment, rather than static pre-scripted settings used here.
5. **Exploring Instinctive Behaviors**: This could be accomplished by differentially "weighting" the connections between certain actions and affective states at the start of training, by heavily weighting them closer to 1. The weights are currently initialized all the same.
6. **Exploring the Effects of Cognitive Dissonance:** This could be defined as when the difference between the predicted next emotion and actual next emotion exceeds some threshold. In other words, when the agent expects the environment to respond in one way, but it instead responds completely differently.
7. **Conscious vs. Automatic Response Mechanisms**: As mentioned in the introduction, the simulation described here could be expanded to explore something akin to behavior-based approaches [8,9], but geared toward toward personality traits. This relates to the ongoing debate Strong vs. Weak AI, and whether overtly engineered systems capable of "conscious thought" are necessary for intelligent behavior [65].
8. **Failure Response:** A critical issue in real-world interactive systems is the ability to respond to "failures" that cause the system to crash, such as some event outside the bounds of expected behavior occurring [74]. For example, this was a major challenge for the robotic face art museum deployment shown in Figure 1 during unconstrained human-robot interaction. Patterns of convergence/deconvergence in social simulations may shed light on this issue.

9. **Approach/Avoidance Behaviors:** As mentioned in Section 4.2, similar simplified models for robot control have been developed for approach/avoidance in a number of studies [70,71]. One potentially interesting alternative to the convergence-based metric we used here would be a performance metric based on exhibited approach/avoidance behaviors relative to some simulated social stimuli.

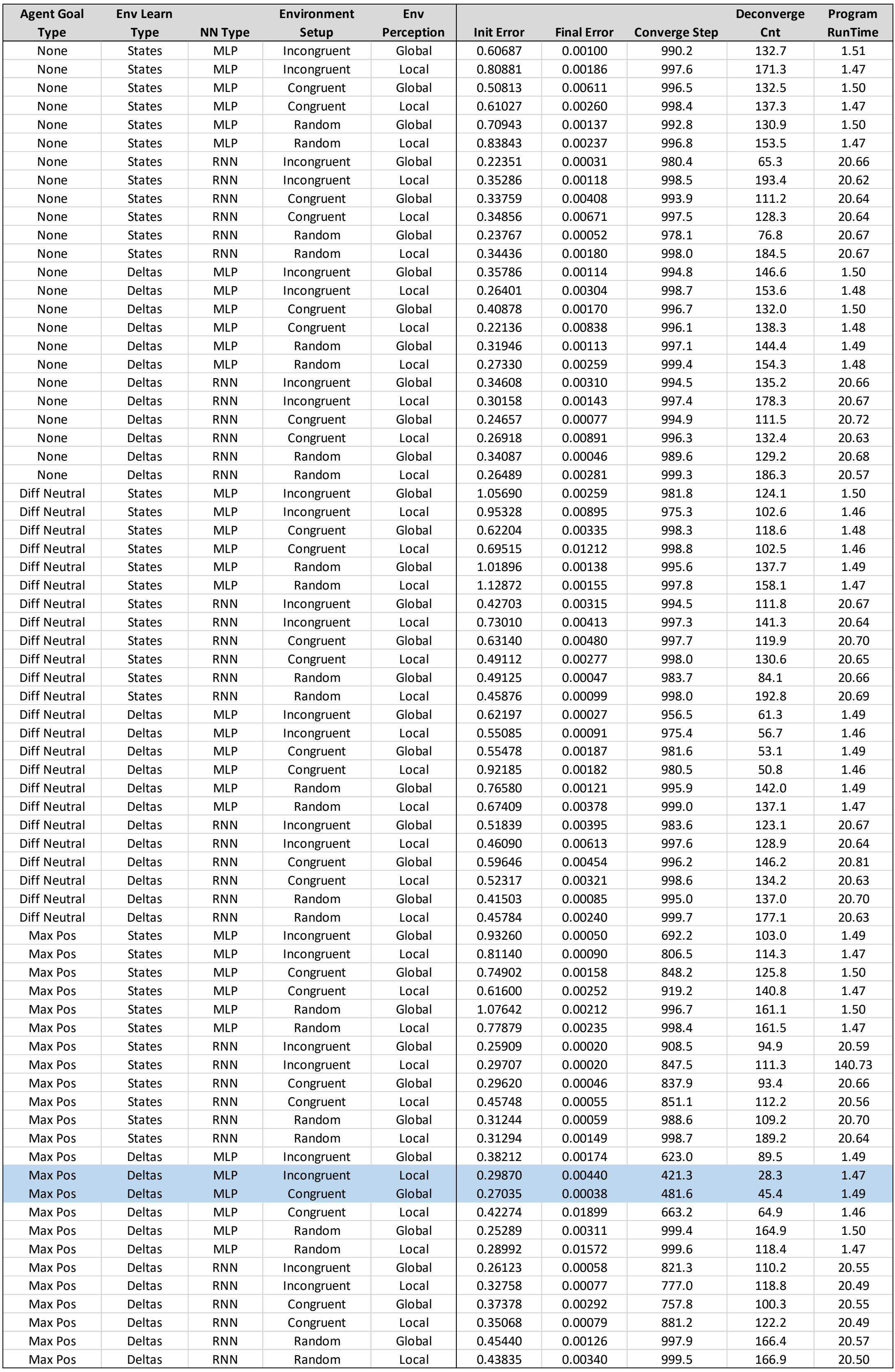
Funding: The data from the physical experiments was funded by NSF grant #IIS-1143712, as well as the NSF/IU IGERT summer research award in cognitive science and Indiana University's School of Informatics, Computing, and Engineering (SICE).

Data Availability Statement: In the spirit of scientific replicability, the Python code for these simulation experiments is made available at the authors' website (REDACTED), to allow others to generate similar data. How the variables in the programming code map to the terms described below is shown in Table 1.

Acknowledgments: The authors would like to thank Sadaf Ebrahimi for their assistance on this work. Additionally, the authors would like to acknowledge their collaborators from the original physical HRI experiments: Selma Sabanovic, Amyra Asamoah, Kay Jessee, Matthew R. Francisco, Kate Shaw, and Marlena Fraune.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Appendix A



Agent Goal Type	Env Learn Type	NN Type	Environment Setup	Env Perception	Init Error	Final Error	Converge Step	Deconverge Cnt	Program RunTime
None	States	MLP	Incongruent	Global	0.60687	0.00100	990.2	132.7	1.51
None	States	MLP	Incongruent	Local	0.80881	0.00186	997.6	171.3	1.47
None	States	MLP	Congruent	Global	0.50813	0.00611	996.5	132.5	1.50
None	States	MLP	Congruent	Local	0.61027	0.00260	998.4	137.3	1.47
None	States	MLP	Random	Global	0.70943	0.00137	992.8	130.9	1.50
None	States	MLP	Random	Local	0.83843	0.00237	996.8	153.5	1.47
None	States	RNN	Incongruent	Global	0.22351	0.00031	980.4	65.3	20.66
None	States	RNN	Incongruent	Local	0.35286	0.00118	998.5	193.4	20.62
None	States	RNN	Congruent	Global	0.33759	0.00408	993.9	111.2	20.64
None	States	RNN	Congruent	Local	0.34856	0.00671	997.5	128.3	20.64
None	States	RNN	Random	Global	0.23767	0.00052	978.1	76.8	20.67
None	States	RNN	Random	Local	0.34436	0.00180	998.0	184.5	20.67
None	Deltas	MLP	Incongruent	Global	0.35786	0.00114	994.8	146.6	1.50
None	Deltas	MLP	Incongruent	Local	0.26401	0.00304	998.7	153.6	1.48
None	Deltas	MLP	Congruent	Global	0.40878	0.00170	996.7	132.0	1.50
None	Deltas	MLP	Congruent	Local	0.22136	0.00838	996.1	138.3	1.48
None	Deltas	MLP	Random	Global	0.31946	0.00113	997.1	144.4	1.49
None	Deltas	MLP	Random	Local	0.27330	0.00259	999.4	154.3	1.48
None	Deltas	RNN	Incongruent	Global	0.34608	0.00310	994.5	135.2	20.66
None	Deltas	RNN	Incongruent	Local	0.30158	0.00143	997.4	178.3	20.67
None	Deltas	RNN	Congruent	Global	0.24657	0.00077	994.9	111.5	20.72
None	Deltas	RNN	Congruent	Local	0.26918	0.00891	996.3	132.4	20.63
None	Deltas	RNN	Random	Global	0.34087	0.00046	989.6	129.2	20.68
None	Deltas	RNN	Random	Local	0.26489	0.00281	999.3	186.3	20.57
Diff Neutral	States	MLP	Incongruent	Global	1.05690	0.00259	981.8	124.1	1.50
Diff Neutral	States	MLP	Incongruent	Local	0.95328	0.00895	975.3	102.6	1.46
Diff Neutral	States	MLP	Congruent	Global	0.62204	0.00335	998.3	118.6	1.48
Diff Neutral	States	MLP	Congruent	Local	0.69515	0.01212	998.8	102.5	1.46
Diff Neutral	States	MLP	Random	Global	1.01896	0.00138	995.6	137.7	1.49
Diff Neutral	States	MLP	Random	Local	1.12872	0.00155	997.8	158.1	1.47
Diff Neutral	States	RNN	Incongruent	Global	0.42703	0.00315	994.5	111.8	20.67
Diff Neutral	States	RNN	Incongruent	Local	0.73010	0.00413	997.3	141.3	20.64
Diff Neutral	States	RNN	Congruent	Global	0.63140	0.00480	997.7	119.9	20.70
Diff Neutral	States	RNN	Congruent	Local	0.49112	0.00277	998.0	130.6	20.65
Diff Neutral	States	RNN	Random	Global	0.49125	0.00047	983.7	84.1	20.66
Diff Neutral	States	RNN	Random	Local	0.45876	0.00099	998.0	192.8	20.69
Diff Neutral	Deltas	MLP	Incongruent	Global	0.62197	0.00027	956.5	61.3	1.49
Diff Neutral	Deltas	MLP	Incongruent	Local	0.55085	0.00091	975.4	56.7	1.46
Diff Neutral	Deltas	MLP	Congruent	Global	0.55478	0.00187	981.6	53.1	1.49
Diff Neutral	Deltas	MLP	Congruent	Local	0.92185	0.00182	980.5	50.8	1.46
Diff Neutral	Deltas	MLP	Random	Global	0.76580	0.00121	995.9	142.0	1.49
Diff Neutral	Deltas	MLP	Random	Local	0.67409	0.00378	999.0	137.1	1.47
Diff Neutral	Deltas	RNN	Incongruent	Global	0.51839	0.00395	983.6	123.1	20.67
Diff Neutral	Deltas	RNN	Incongruent	Local	0.46090	0.00613	997.6	128.9	20.64
Diff Neutral	Deltas	RNN	Congruent	Global	0.59646	0.00454	996.2	146.2	20.81
Diff Neutral	Deltas	RNN	Congruent	Local	0.52317	0.00321	998.6	134.2	20.63
Diff Neutral	Deltas	RNN	Random	Global	0.41503	0.00085	995.0	137.0	20.70
Diff Neutral	Deltas	RNN	Random	Local	0.45784	0.00240	999.7	177.1	20.63
Max Pos	States	MLP	Incongruent	Global	0.93260	0.00050	692.2	103.0	1.49
Max Pos	States	MLP	Incongruent	Local	0.81140	0.00090	806.5	114.3	1.47
Max Pos	States	MLP	Congruent	Global	0.74902	0.00158	848.2	125.8	1.50
Max Pos	States	MLP	Congruent	Local	0.61600	0.00252	919.2	140.8	1.47
Max Pos	States	MLP	Random	Global	1.07642	0.00212	996.7	161.1	1.50
Max Pos	States	MLP	Random	Local	0.77879	0.00235	998.4	161.5	1.47
Max Pos	States	RNN	Incongruent	Global	0.25909	0.00020	908.5	94.9	20.59
Max Pos	States	RNN	Incongruent	Local	0.29707	0.00020	847.5	111.3	140.73
Max Pos	States	RNN	Congruent	Global	0.29620	0.00046	837.9	93.4	20.66
Max Pos	States	RNN	Congruent	Local	0.45748	0.00055	851.1	112.2	20.56
Max Pos	States	RNN	Random	Global	0.31244	0.00059	988.6	109.2	20.70
Max Pos	States	RNN	Random	Local	0.31294	0.00149	998.7	189.2	20.64
Max Pos	Deltas	MLP	Incongruent	Global	0.38212	0.00174	623.0	89.5	1.49
Max Pos	Deltas	MLP	Incongruent	Local	0.29870	0.00440	421.3	28.3	1.47
Max Pos	Deltas	MLP	Congruent	Global	0.27035	0.00038	481.6	45.4	1.49
Max Pos	Deltas	MLP	Congruent	Local	0.42274	0.01899	663.2	64.9	1.46
Max Pos	Deltas	MLP	Random	Global	0.25289	0.00311	999.4	164.9	1.50
Max Pos	Deltas	MLP	Random	Local	0.28992	0.01572	999.6	118.4	1.47
Max Pos	Deltas	RNN	Incongruent	Global	0.26123	0.00058	821.3	110.2	20.55
Max Pos	Deltas	RNN	Incongruent	Local	0.32758	0.00077	777.0	118.8	20.49
Max Pos	Deltas	RNN	Congruent	Global	0.37378	0.00292	757.8	100.3	20.55
Max Pos	Deltas	RNN	Congruent	Local	0.35068	0.00079	881.2	122.2	20.49
Max Pos	Deltas	RNN	Random	Global	0.45440	0.00126	997.9	166.4	20.57
Max Pos	Deltas	RNN	Random	Local	0.43835	0.00340	999.5	166.9	20.50

Table A1: Main Results (Full Table)

References

1. Dexter, J.P.; Prabakaran, S.; Gunawardena, J. A Complex Hierarchy of Avoidance Behaviors in a Single-Cell Eukaryote." *Current Biology*, 2019, 29(24), 4323-4329.
2. Robert, L. Personality in the human robot interaction literature: A review and brief critique. *Proceedings of the 24th Americas Conference on Information Systems,* 2018, pp. 16-18.
3. Lisetti, C.; Hudlicka, E. Why and how to build emotion-based agent architectures. In: *The Oxford Handbook of Affective Computing*, Oxford University Press, USA, 2015, pp. 94.
4. Bennett, C.C.; Šabanović, S. Deriving minimal features for human-like facial expressions in robotic faces. *International Journal of Social Robotics*, 2014, 6(3), 367-381.
5. Bennett, C.C.; Šabanović, S. The effects of culture and context on perceptions of robotic facial expressions. *Interaction Studies*, 2015, 16(2), 272-302.
6. Bennett, C.C. *Robotic faces: Exploring dynamical patterns of social interaction between humans and robots*. Indiana University Press, USA, 2015.
7. Hönemann, A.; Bennett, C.C; Wagner, P.; Sabanovic, S. Audio-visual synthesized attitudes presented by the German speaking robot SMiRAE. *Proceedings of the 15th International Conference on Auditory-Visual Speech Processing*. 2015.
8. Brooks, R.A. From earwigs to humans. *Robotics and Autonomous Systems*. 1997, 20(2-4), 291-304.
9. Sun, R. (Ed.) *Cognition and Multi-Agent Interaction: From Cognitive Modeling to Social Simulation*. Cambridge University Press, UK, 2006.
10. Masterson, J.F. *The Search for the Real Self: Unmasking the Personality Disorders of Our Age*. Taylor & Francis, New York, NY, USA, 1988.
11. Lieb, K.; Zanarini, M.C.; Schmahl, C.; Linehan, M.M.; Bohus, M. Borderline personality disorder. *The Lancet*, 2004, 364(9432), 453-461.
12. Ortony, A.; Clore, G. L.; Collins, A. *The Cognitive Structure of Emotions*. Cambridge University Press, UK, 1988
13. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Computation*, 1997, 9(8), 1735-1780.
14. Scherer, KR.; Schorr, A.;Johnstone, T. (Eds.). *Appraisal Processes in Emotion: Theory, Methods, Research*. Oxford University Press, UK, 2001.
15. Murphy, R.R.; Lisetti, C.L.; Tardif, R.; Irish, L.; Gage, A. Emotion-based control of cooperating heterogeneous mobile robots. *IEEE Transactions on Robotics and Automation*, 2002, 18(5), 744-757.
16. Tonkin, M.; Vitale, J.; Herse, S.; Williams, M.A.; Judge, W.; Wang, X. Design methodology for the UX of HRI: A field study of a commercial social robot at an airport. *ACM/IEEE International Conference on Human-Robot Interaction (HRI),* 2018, pp. 407-415.
17. Cruz-Maya, A.; Tapus, A. Influence of user's personality on task execution when reminded by a robot. *International Conference on Social Robotics (ICSR),* 2016, pp. 829-838.
18. Gockley, R.; Mataric, M.J. Encouraging physical therapy compliance with a hands-off mobile robot. *ACM/IEEE International Conference on Human-Robot Interaction (HRI),* 2006, pp. 150-155.
19. Kimoto, M.; Iio, T.; Shiomi, M.; Tanev, I.; Shimohara, K.; Hagita, N. Relationship between personality and robots' interaction strategies in object reference conversations. *Proceedings of the Second International Conference on Electronics and Software Science (ICESS2016),* 2016, pp. 128-136.
20. Lee, K.M.; Peng, W.; Jin, S.A.; Yan, C. Can robots manifest personality?: An empirical test of personality recognition, social responses, and social presence in human–robot interaction. *Journal of Communication*, 2006, 56(4), 754-772.
21. Looije, R.; Neerincx, M.A.; Cnossen, F. Persuasive robotic assistant for health self-management of older adults: Design and evaluation of social behaviors. *International Journal of Human-Computer Studies*, 2010, 68(6), 386-397.
22. Park, E.; Jin, D.; Del Pobil, A.P. The law of attraction in human-robot interaction. *International Journal of Advanced Robotic Systems*, 2012, 9(2), 35.
23. Sandoval, E.B.; Brandstetter, J.; Obaid, M.; Bartneck, C. Reciprocity in human-robot interaction: a quantitative approach through the prisoner's dilemma and the ultimatum game. *International Journal of Social Robotics,* (2016) 8(2), 303-317.
24. Bartneck D.; Kulic E.; Croft M.; Zoghbi, S. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics,* 2009, 1, 71-81.
25. Santamaria, T.; Nathan-Roberts, D. Personality Measurement and Design in Human-Robot Interaction: A Systematic and Critical Review. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting,* 2017, 61(1), 853-857.
26. Grollman, D. H. Avoiding the Content Treadmill for Robot Personalities. *International Journal of Social Robotics*, 2018, 10(2), 225-234.
27. Ekman P.; Friesen W.V. *Unmasking the Face: A Guide to Recognizing Emotions from Facial Clues.* Malor Books, Los Altos, CA, USA, 2003.
28. Russell JA; Fernández-Dols J.M.. *The Psychology of Facial Expression*. Cambridge University Press, Cambridge, UK, 1997.
29. Breazeal, C. Emotion and sociable humanoid robots. *International Journal of Human-Computer Studies*, 2003, 59(1-2), 119-155.

30. Han, M.J.; Lin, C.H.; Song, K.T. Robotic emotional expression generation based on mood transition and personality model. *IEEE Transactions on Cybernetics*, 2012, 43(4), 1290-1303.
31. Ilies, R.; Judge, T.A. Understanding the dynamic relationships among personality, mood, and job satisfaction: A field experience sampling study. *Organizational Behavior and Human Decision Processes*, 2002, 89(2), 1119-1139.
32. Meyer, G.J.; Shack, J.R. Structural convergence of mood and personality: Evidence for old and new directions. *Journal of Personality and Social Psychology*, 1989, 57(4), 691.
33. Van Steenbergen, H.; Band, G.P.; Hommel, B. In the mood for adaptation: How affect regulates conflict-driven control. *Psychological Science*, 2010, 21(11), 1629-1634.
34. Ivanović, M.; Budimac, Z.; Radovanović, M.; Kurbalija, V.; Dai, W.; Bădică, C.; et al. Emotional agents–state of the art and applications. *Computer Science and Information Systems*, 2015, 12(4), 1121-1148.
35. Bosse, T.; Duell, R.; Memon, Z.A.; Treur, J.; van der Wal, C.N. Agent-based modeling of emotion contagion in groups. *Cognitive Computation*, 2015, 7(1), 111-136.
36. Evers, E.; de Vries, H.; Spruijt, B.M.; Sterck, E.H. The EMO-model: an agent-based model of primate social behavior regulated by two emotional dimensions, anxiety-FEAR and satisfaction-LIKE. PloS One, 2014 9(2), e87955.
37. Shen, Z.; Miao, C. An emotional agent in virtual learning environment. In: *Transactions on Edutainment IV*. Springer, Berlin, Germany, 2010, pp. 22-33.
38. Puică, M.A.; Florea, A.M. Emotional belief-desire-intention agent model: Previous work and proposed architecture. *International Journal of Advanced Research in Artificial Intelligence*, 2013, 2(2), 1-8.
39. Bosse, T.; De Lange, F.P. Development of virtual agents with a theory of emotion regulation. *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, 2008, pp.461-468.
40. Gratch, J. The Social Psychology of Human-agent Interaction. *Proceedings of the 7th International Conference on Human-Agent Interaction (HAI)*, 2019, pp. 1-1.
41. MacDorman, K.F.; Chattopadhyay, D. Reducing consistency in human realism increases the uncanny valley effect; increasing category uncertainty does not. *Cognition*, 2016, 146, 190-205.
42. Dang, T.H.H.; Hutzler, G.; Hoppenot, P. Emotion modeling for intelligent agents-Towards a unifying framework. *IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, 2011, 3, 70-73.
43. Laird, J. E. *The Soar Cognitive Architecture*. MIT press, Cambridge MA, USA, 2012.
44. Picard, R. W. *Affective Computing*. MIT press Cambridge MA, USA, 2000.
45. Šabanović, S. Robots in society, society in robots. *International Journal of Social Robotics*, 2010, 2(4), 439-450.
46. Sabanovic, S.; Michalowski, M.P.; Simmons, R. Robots in the wild: Observing human-robot social interaction outside the lab. *IEEE International Workshop on Advanced Motion Control*, 2006, pp. 596-601.
47. Jung, M.; Hinds, P. Robots in the wild: A time for more robust theories of human-robot interaction. *ACM Transactions on Human-Robot Interaction (THRI)*, 2018, 7(1), 2.
48. Cass, A.G.; Striegnitz, K.; Webb, N.; Yu, V. Exposing real-world challenges using HRI in the wild. *Proceedings of the 4th Workshop on Public Space Human-Robot Interaction at the Intl. Conf. on Human-Computer Interaction with Mobile Devices and Services*, 2018.
49. Šabanović, S.; Bennett, C.C.; Lee, H. R. Towards culturally robust robots: A critical social perspective on robotics and culture. *Proceedings of the HRI Workshop on Culture-Aware Robotics*, 2014.
50. Schneider, D.; Lam, R.; Bayliss, A.P.; Dux, P.E. Cognitive load disrupts implicit theory-of-mind processing. *Psychological Science*, 2012, 23(8), 842-847.
51. Spunt, R.P.; Lieberman, M.D. The busy social brain: evidence for automaticity and control in the neural systems supporting social cognition and action understanding. *Psychological Science*, 2013, 24(1), 80-86.
52. Paas, F.; Van Gog, T.; Sweller, J. Cognitive load theory: New conceptualizations, specifications, and integrated research perspectives. *Educational Psychology Review*, 2010, 22(2), 115-121.
53. Smith, S.M.; Petty, R.E. Personality moderators of mood congruency effects on cognition: The role of self-esteem and negative mood regulation. *Journal of Personality and Social Psychology*, 1995, 68(6), 1092.
54. Robins, B.; Dautenhahn, K.; Te Boekhorst, R; Nehaniv, C.L. Behaviour delay and robot expressiveness in child-robot interactions: a user study on interaction kinesics. *ACM/IEEE international conference on Human robot Iinteraction (HRI*), 2008 pp. 17-24.
55. Barsalou, L.W.; Breazeal, C.; Smith, L.B. Cognition as coordinated non-cognition. *Cognitive Processing*, 2007, 8(2), 79-91.
56. Beer, R.D. Dynamical approaches to cognitive science. *Trends in Cognitive Sciences*, 2000, 4(3), 91-99.
57. Vallacher, R.R.; Read, S.J.; Nowak, A. The dynamical perspective in personality and social psychology. *Personality and Social Psychology Review*, 2002, 6(4), 264-273.
58. Kastner, S.; De Weerd, P.; Desimone, R.; Ungerleider, L.G. Mechanisms of directed attention in the human extrastriate cortex as revealed by functional MRI. *Science*, 1998, 282(5386), 108-111.
59. Land M.F.; Nilsson D.E. *Animal Eyes*. Oxford University Press, Oxford, UK, 2012.
60. Movellan J.R.; Tanaka F.; Fortenberry B.; Aisaka K. The RUBI/QRIO project: Origins, principles, and first steps. *IEEE International Conference on Development and Learning (ICDL)*, 2005, pp.80-86
61. Bennett, C.C.; Hauser, K. Artificial intelligence framework for simulating clinical decision-making: A Markov decision process approach. *Artificial Intelligence in Medicine*, 2013, 57(1), 9-19.
62. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Computation*, 1997, 9(8), 1735-1780.

63. Cover T.M; Thomas J.A. *Elements of Information Theory*. John Wiley and Sons , Hoboken, NJ, 2006.
64. Shannon, C.E. The mathematical theory of communication. *The Bell System Technical Journal*, 1948, 27, 379–423.
65. Clark, A. Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain sciences*, 2013, 36(3), 181-204.
66. Hosoya, T.; Baccus, S.A.; Meister, M. Dynamic predictive coding by the retina. *Nature*, 2005, 436(7047), 71.
67. Broz, F.; Nehaniv, C.L.; Kose, H.; Dautenhahn, K. Interaction Histories and Short-Term Memory: Enactive Development of Turn-Taking Behaviours in a Childlike Humanoid Robot. *Philosophies*, 2019, 4(2), 26.
68. Bashir, F. I.; Khokhar, A.A.; Schonfeld, D. Object trajectory-based activity classification and recognition using hidden Markov models. *IEEE Transactions on Image Processing*, 2007, 16(7), 1912-1919.
69. Xia, L.; Chen, C.C.; Aggarwal, J.K. View invariant human action recognition using histograms of 3D joints. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 20-27.
70. Lones, J.; Lewis, M.; Cañamero, L. Hormonal modulation of interaction between autonomous agents. *4th International Conference on Development and Learning and on Epigenetic Robotics*, 2014, pp. 402-407.
71. Belkaid, M.; Cuperlier, N.; Gaussier, P. Emotional modulation of peripersonal space as a way to represent reachable and comfort areas. *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2015, pp. 353-359.
72. Grant B.F.; Chou S.P.; Goldstein R.B.; Huang B.; Stinson F.S.; Saha T.D.; et al. Prevalence, correlates, disability, and comorbidity of DSM-IV Borderline Personality Disorder: Results from the Wave 2 National Epidemiologic Survey on Alcohol and Related Conditions. *Journal of Clinical Psychiatry*. 2008, 69(4): 533–545.
73. Fonagy P. Attachment and borderline personality disorder. *Journal of the American Psychoanalytic Association*, 2000, 48(4): 1129-1146.
74. Honi, S.; Oron-Gilad, T. Understanding and resolving failures in human-robot interaction: Literature review and model development. *Frontiers in Psychology*, 2018, 9, 861.