Deliberative and Affective Reasoning: a Bayesian Dual-Process Model

Jesse Hoey
Cheriton School of Computer Science
University of Waterloo
Waterloo, Ontario, Canada
jhoey@cs.uwaterloo.ca

Zahra Sheihkbahaee
Cheriton School of Computer Science
University of Waterloo
Waterloo, Ontario, Canada
zsheikb@uwaterloo.ca

Neil J. MacKinnon
University of Guelph
Guelph, Ontario, Canada
nmackinn@uoguelph.ca

***Abstract*—The presence of artificial agents in human social networks is growing. From chatbots to robots, human experience in the developed world is moving towards a socio-technical system in which agents can be technological or biological, with increasingly blurred distinctions between. Given that emotion is a key element of human interaction, enabling artificial agents with the ability to reason about affect is a key stepping stone towards a future in which technological agents and humans can work together. This paper presents work on building intelligent computational agents that integrate both emotion and cognition. These agents are grounded in the well-established social-psychological Bayesian Affect Control Theory (*BayesAct*). The core idea of *BayesAct* is that humans are motivated in their social interactions by affective alignment: they strive for their social experiences to be coherent at a deep, emotional level with their sense of identity and general world views as constructed through culturally shared symbols. This affective alignment creates cohesive bonds between group members, and is instrumental for collaborations to solidify as relational group commitments. *BayesAct* agents are motivated in their social interactions by a combination of affective alignment and decision theoretic reasoning, trading the two off as a function of the uncertainty or unpredictability of the situation. This paper provides a high-level view of dual process theories and advances *BayesAct* as a plausible, computationally tractable model based in social-psychological and sociological theory.**

***Index Terms*—affect control theory, Markov decision process, free energy, active inference, social order, artificial intelligence**

I. Introduction

A key element of human experience is emotion, and enabling artificial agents with the ability to reason about emotions is a key stepping stone towards a future in which artificial intelligence (AI) and humans can work together cooperatively in social dilemmas[1], while respecting ethical, moral and normative orders in society. Our vision is to build intelligent computational agents that parsimoniously integrate both emotion and cognition, that are able to become members of a socio-technical system. We ground our vision in a social-psychological theory of affective alignment and social order called *BayesAct* [19], [33], which is based on the sociological Affect Control Theory [17], [26].

Funded by the Canadian Natural Sciences and Engineering Research Council and Social Sciences and Humanities Research Council.

[1]A social dilemma is a game with *uncompensated interdependencies* (externalities) [24]: each person's actions in the game affect other persons without their explicit consent (e.g. without compensating them).

A key area of application for emotionally aligned AI agents is collaborative networks. Understanding the social forces behind self-organized collaboration is increasingly important in today's society. More than ever, technological and social innovations are enabled by information and communication technologies and are generated through informal, distributed processes of collaboration, rather than in formal, hierarchical or market-based organizations. Although an individualization narrative pervades much theorizing about twenty-first century human interactions, an alternative socio-relational narrative has recently developed in which relational and affective person-to-group ties are understood as a keystone of networked coordination and effectiveness [25]. Relational ties grow from repeated interactions in groups with a shared responsibility in which positive emotions are created. Attribution by group members of their feelings (affect) to the group further strengthens the relational ties, creating a self-reinforcing mechanism for group coordination. Shared responsibility and positive affective interactions make the group salient and endow it with a moral and normative force upon the group members. Groups thus endowed are powerful agents for the mobilization of collaborative human efforts and collective action.

In this paper, we first discuss dual process models in general in Section II, and *BayesAct* in particular in Section III. In Section V, we review two key application areas, one in online collaborative networks (e.g. GitHub) [20] and the other in building assistive technologies for persons with dementia [32]. We briefly discuss ethics in Section VI and conclude. Technical details about *BayesAct* can be found at `bayesact.ca`.

II. Dual Process Models

Human and artificial intelligent agents are faced with the computational problem posed by the complexity of inputs to the senses. Agents must find a way to map this high dimensional input space into an equally high dimensional action space. Agents can handle the complexity of the input by constructing a representation of it, and then preforming calculations over this representation. We will call this the *denotative* representation, and it is an abstraction of the physics of environment. For example, it is able to represent the positions of pieces on a chess board and make predictions about how a game will turn out given a sequence of moves,

or it can represent the bids in a negotation. The denotative representation is assumed to be symbolic, but can be implemented sub-symbolically [2]. However, denotative calculations rapidly become intractable, and are exacerbated when the environment includes other intelligent agents [12].

As the complexity of the environment increases, an agent with fixed computational resources runs into a bound that prevents it from modeling the added complexity. This resource bound results in a more dispersed estimates of the denotative state representation because the predictions made by the agent's (limited) internal denotative representation have increasing trouble matching the evidence from the world. Such an agent can handle this inability to predict the future by believing the predictions and ignoring the evidence (underfitting), or by relaxing the predictions and believing the evidence (overfitting). While the first agent will have difficulty adapting to change (but can see it coming), the second will have difficulty predicting change (but can adapt to it).

An agent with a hierarchical model can do both, however, as it can create and track a lower-dimensional version of the denotative state that allows it to continue making predictions, albeit with reduced precision. In machine learning, the basic idea of approximating a complex function with another, simpler, function is referred to as variational, and if the functions are probabilistic, then variational Bayesian. An intelligent agent that is using a variational optimization technique to optimize policies of behaviour can be seen as an instance of active inference. Active Inference [15] proposes consistency between an agent's internal model and the environment in which it is embedded as a fundamental principle underlying biological agents. The complexity of the agent's environment, defined as the total number of configurations accessible to the agent, is the true *free energy*. As an agent's world becomes more complex (e.g. with the addition of other social agents), this true free energy becomes intractable to model within an agent's resource bound, and so the agent must approximate. An agent's *variational free energy* is an internal measure of how well its (approximate) model fits the real world. The idea of minimizing this variational free energy is the same as moving towards a state of true free energy which is minimal and is consistent with the world in which it is embedded [14].

BayesAct is a computational dual process (hierarchical) model of intelligence, in which one process is continuous in one or more dimensions and is equated with human sentiment, while the other process is discrete or continuous and models human deliberative reasoning and decision making [19], [33]. The dual process is built to handle uncertainty and surprise, naturally shifting between higher bias (lower variance) models in the sentiment space in more (denotatively) uncertain (invalid/unpredictable) situations, to higher precision (lower bias) models in the deliberative space in more predictable (valid/predictable) environments. In *BayesAct*, we avoid the terms "cognitive" and "emotional", and refer instead to a *"denotative"* representation and a *"connotative"* one. The "connotative" is also referred to as "sentiment" or "feeling", both of which are "affective", while "emotion" refers specifically to a signaling mechanism described in Section III. The denotative representation requires deliberative reasoning, in which sequences of futures are examined in memory to allow for selection of appropriate actions in the present. The connotative representation, on the other hand, is the *meaning* of the world at the level of sentiments or feelings in a relatively low dimensional space, and produces indications of (in)consistency or (in)coherence. Direct evidence of these meanings is obtained through emotional signals from other humans. Importantly, the consistency encoded in sentiments extends to agent actions and provides a rough (heuristic) guide over policies. The social intelligence provided by this consistency is shared by agents in a community, and motivates them to want to do things according to the same practice ("habitus" [5]) which encodes the "way we do things". This shared practice is an approximation built to handle and alleviate computational complexity of the social world.

Consider a simple example in which we characterize people and behaviours as either "good" or "bad". Cultural consensus is that good people will do good things, and so the denotative state does not have to model the situation in which good people are doing bad things, and thus can be simpler. The connotative state can therefore be linked to the denotative state with some energy functional dependent on the discrepancy between the meanings or feelings of things out of context and in context. In our example, we know there are "good" people and "good" behaviours (out of context), and we expect these to be found together (in context). If we find inconsistency (good people doing bad things), then this will be surpising, cause increased dispersion in estimates at the connotative level, and pushing reasoning into the denotative level for analysis and re-labeling of actors and behaviours (maybe this is not, actually a "good" person, or maybe the behaviour is not a "bad" behaviour?).

In *BayesAct*, we associate the connotative state as a variational approximation to the denotative state, and identify it as a mechanism for encoding efficient policies in the environment that includes a (potential) social group. Agents minimizing their free energy using such a dual-process model will engage in active inference and will learn a model of emotional dynamics that can ease the computational load on the denotative representation. Further, if the variational approximations are linked across agents, then the resulting social group will learn to *share* the same approximation in order to more efficiently solve social dilemmas. It is precisely because the collaborative solution to the dilemma yields higher payouts for all individuals that the connotative representation is selected because it is consistent with that of other agents. A secondary, normally multi-modal, signaling mechanism is used to facilitate this linkage across agents. These signals are termed "emotional" and define an "emotion language" [36] that ensures that the agents' connotative spaces are directly linked. This emotion

[2] By subsymbolic, we mean as in a neural network, where the "symbols" are weights on neurons, and therefore somewhat difficult to interpret. However, we do not rule out the possibility that a subsymbolic representation could be used to model a symbolic one. For example in a deep reinforcement learning problem, the symbols are the actions being predicted (in fact, the values of these actions), while the neural network simply provides the mapping function.

language, communicated in part through facial expressions and paralinguistics, allows agents to communicate what aspects of the denotative state are worth more fully exploring. As a result, the connotative state acts as a "flashlight" that illuminates the *same* part of the denotative state for all agents who share it. Such "splotlight" methaphors have been deeply explored in the context of psychological (usually visual) attention [8]. Once this linking occurs, then the connotative state encodes a *social contract* with the social group in which the agent finds itself.

Dual process theories are well studied in social psychology [7], but many different terms are used to refer to the two levels of processing. The denotative level is often referred to as deliberative, reflective or "System 2", whereas the connotative level as emotional, automatic, or "System 1". In many dual process theories (e.g. [12]), both deliberative and automatic systems are modeled denotatively in a constraint satisfaction type of network. In *BayesAct*, the connotative level is *affective* and serves as a low-dimensional approximation. However, the connotative level is a cognitive process, and so not at the level of "primary" or reactive emotions (e.g. startle reflexes), or of core affect [4], but rather at the level of routine or reflective interpretations of emotions linked to procedural memory [29].

Behavioural economists have also tackled emotional human motivations, usually by proposing that humans make choices based on a modified utility function that includes some reward for fairness [31] or penalty for inequity [11] or conformity [27]. However, heuristic adjustments may not be comprehensive enough to account for human behaviour across all situations, and a morality concept that is not based on outcomes can be used as a more parsimonious account [6]. The question of how this morality is defined is left open.

Athough it is increasingly clear that humans operate with something akin to a dual process model [38], some (e.g. artificial intelligence practitioners) may argue that a connotative representation is unecessary for general intelligence, and that sufficient resources (relaxing the resource bound) will lead to fully denotative, decision-theoretically rational agents, or "econs". The higher precision allowed by a denotative representation seems to point the way to a superintelligence [21], and a decision-theoretically rational social system. We propose that artificial intelligence *requires* a dual process denotative/connotative model to exist as a general-purpose member of a socio-technical system. We present computational, evolutionary and social arguments here.

One argument is that an agent's ability to model the vast numbers of combinations of other agents becomes challenging unless a lower dimensional manifold is discovered that enables cooperation in groups. Consider a multi-agent system consisting of agents of N different types who can behave in N_B different ways. If each agent attempts to model all other agents in its group, including their first-level (direct) interactions with each other, the number of combinations would be factorial in the product $N \times N_B$. If $N = 150$ [9] and N_B is 500 or so, then representation is intractable [3].

[3]Indeed, with only $N = M = 10$, the number of combinations is astronomical, approaching the number of atoms in the universe.

The second argument in support of a connotative state is offered by Turner [36] from an evolutionary perspective, who notes that early apes were forced into the forest canopy by other simians around 25 million years ago, and had to deal with a more complex, three dimensional space, making permanent groups more difficult and leading to a species with no permanent bonds and increased promiscuity. This increased complexity entails an increase in the true free energy (number of configurations the world can be in) that the apes had to model, and pressured the development of approximations to this increased free energy. When the descendents of these apes, the early hominids, were forced into the savannah in an era of climate change around 10 million years ago, the reduced complexity of a two dimensional world, combined with the need for stronger group cohesion (because of predators) pushed these approximations to other uses, fostered the development of early "emotional" languages, and allowed larger structures of humans to form, opening the door for collective activities like solving social dilemmas. From an evolutionary perspective, the perserverance of this emotional language and related structures is an indication of their usefulness in the context of human groups, and therefore we expect it to be useful in a group involving artificial agents as well.

The final argument is simple. A group of agents who are able to coordinate to solve social dilemmas will be more suited for survival than a group that does not. This coordination departs from the principles of decision-theoretic (individual) rationality, but can be enforced by a connotative representation that inextricably links agents through emotional signaling. This inextricable link provides a mechanism for a group of agents to jointly minimize their free energy in an uncertain world, and is thus a provable statement.

III. *BayesAct*

We present here a short introduction to ACT and *BayesAct*, and refer the reader to longer treatments in [19], [33], covering relationships to other theories of emotion (e.g. appraisal).

Affect Control Theory (ACT) [17], [26] proposes a fundamental link between symbolic, denotative, representations of the social environment and the continuous, connotative, representations of the sentiments or feelings associated with those denotative representations. For example, when one perceives a person in a white coat in a hospital, a denotative impression is formed of this person that is represented with a symbol (doctor). This symbol has an associated "fundamental" sentiment in a three-dimensional affective "EPA" space of evaluation (E: good/bad), power (P: strong/weak) and activity (A: active/inactive). Doctors, for example, usually evoke feelings of goodness, strength, and modest activity. EPA space has been found through decades of research to be a cross-culturally normative representation of meaning [30].

The link between denotative and connotative in ACT is empirically determined through population surveys using semantic differential scales. These measurements yield a set of samples from a population distribution in the sentiment space, which can then be parametrically estimated (e.g. as the mean and variance of a normal distribution), or non-parametrically

represented (as a set of samples). Lists of such measurements are called "dictionaries" of mappings from labels to sentiment. In ACT, only the mean of this measurement is used to link denotative and connotative. Thus, a *doctor* is represented connotatively as (EPA:$\{2.7, 3.0, 0.2\}$) [4]. Given a connotative (EPA) vector, a denotative label can be assigned in ACT using a simple nearest neighbour method (e.g. the closest label to (EPA:$\{-1.0, 2.0, 2.0\}$) is *politician* (EPA:$\{-0.9, 2.3, 1.5\}$) - at a Euclidean distance of 0.35). ACT proposes that events in the world, interpreted symbolically (denotatively), create re-assessments at the connotative level called *transient impressions* that are used to motivate agents towards behaviours that reduce the incoherence between in-context impressions and out-of-context sentiments. This motivation to socially conforming actions can be interpreted as an instance of Bourdieu's "habitus" [5], as explored in more detail in [1].

Emotions in *ACT* are defined precisely as the vector difference between fundamental (out-of-context) and transient (in-context) sentiments, and are a mechanism to help agents signal (in)coherence to each other (e.g. with facial expressions or paralinguistics). Importantly, these signals are not scalar indications of (in)coherence, but rather vector signals giving recipies for restorative behaviour and emotion regulation [16]. For example, if a doctor "talks down to" (EPA:$\{-1.6, -0.1, 0.3\}$) another doctor, the object agent is made to feel less powerful (drops to -0.1) than expected, and will display exasperation or indigance. Upon receiving this signal, the acting agent may restore fundamental sentiments by "making up with" the other.

BayesAct [19], [33] generalises ACT by explicitly representing the distribution over sentiment in a two-level partially observable Markov decision process (POMDP). *BayesAct* models individual differences as variance in sentiments, and modulates the predictions of ACT due to the differences between denotative entities with low and high connotative variances [13]. In the original formulation of the *BayesAct* model, the sentiment was directly observed in an interaction as a three-dimensional, continuous vector that gave a direct measurement of the sentiment of the behaviour being performed. That is, if a doctor was observed injecting someone with medicine, then *BayesAct* expected a direct observation of the mean EPA rating for that denotative behaviour, *inject someone with medicine*: (EPA:$\{0.9, 1.7, -0.2\}$). *BayesAct* had a denotative state, but this only represented elements of an interaction outside of the social definition of identities, such as the state of a game being played. For example, this might be the positions of both agents' pieces on a chessboard, or current bids in a negotiation.

However, the *BayesAct* model can include a denotative representation of identities and behaviours of other agents. If it does so, then these denotative elements are linked to the connotative state through a potential function that measures the incoherence (difference) between the current estimate of the denotative state (e.g. *doctor*) and the current estimate of the connotative state (a distribution in the affective EPA space). For example, if the doctor performs some behaviour uncharacteristic of a doctor (e.g., *abuse* a patient), this doctor would seem less good (lower E) than the culturally accepted definition of a doctor. The incoherence generated between the out-of-context sentiment about doctors (high E) and the impressions created by the observed behaviour pushes the observing agent to a higher energy state. While behaviours can be selected (as in ACT) to reduce incoherence (and thus energy), the energy function can also be used to probabilistically rank likely identities that could be used for re-identification. Thus, if a doctor is observed *harassing* (EPA:$\{-3.0, 0.6, 1.6\}$) a patient (EPA:$\{0.6, -1.5, -1.3\}$), agents would be motivated to act in such a way as to stop the behaviour, or would be forced to re-interpret the doctor as some other identity (the optimal in this case would be (EPA:$\{-4.3, 1.4, 1.7\}$), with a closest label of *rapist* at a distance of 0.44).

Finally, *BayesAct* has two sets of observations. One represents signals about the environment giving evidence for the denotative state. The other represents emotional signals from other agents, and gives direct evidence for the connotative state. Information flows into the model from both connotative and denotative sides, and *BayesAct* computes posterior distributions that best merge the two in a Bayesian sense. Emotion signals are crucial for grounding the connotative state, as otherwise it could be arbitrarily transformed between agents and would be harder to learn.

IV. Uncertainty, Information, and Affect

The link in *BayesAct* between denotative and connotative induces a natural (Bayesian) tradeoff due to relative uncertainty. As the environment becomes less valid (less predictable or more uncertain [23], so the distribution over denotative states is more dispersed or has higher entropy), then the posterior will be more heavily influenced by the prior in the connotative state. Agents in less valid (less predictable) environments will put more weight on the connotative representation: they will make inferences and choose actions that are more in line with connotative (socio-cultural) expectations. In more valid environments, a lower entropy denotative distribution dominates the posterior. Agents in more valid environments will thus act more in line with denotative states and predictive dynamics, and so will be information seekers and utilizers. In a social dilemma, for example, one would expect the agents in less valid environments to cooperate (act according to social prescriptions), while agents in more valid environments will defect (act decision theoretically rationally). This is in line with experiments showing how humans tend to act more prosocially (cooperate in a public goods game) in ambiguous situations (ones in which risk is hard to evaluate, see [37]). In *BayesAct*, risk is represented by the transition dynamics parameters in the denotative space. If the distribution over these parameters has lower entropy, then risk is more well defined, and so ambiguity (the uncertainty in risk) is lower.

In *BayesAct*, any denotative state can be mapped into the same connotative space, allowing for comparisons between actions and identities, for example. The connotative state is required to guide an agent towards socially acceptable choices

[4]For historical reasons, EPA measurements are scaled to lie between -4.3 and +4.3. All data in this paper is taken from a survey of 1742 people in the USA in 2015, see https://research.franklin.uga.edu/act/.

of behavior that can ensure more globally optimal solutions to social dilemmas. Note that this is a different concept than Simon's bounded rationality (SBR) [34]. In SBR, the agent first performs an analysis at the symbolic level (denotative), and then freezes this analysis into a second denotative space called habits and coping. In ACT and in *BayesAct*, the agent gathers a fast impression and then makes predictions in an emotional space with a simple predictive function which can rapidly generate somewhat (socially) relevant predictions about future outcomes involving other agents. In *BayesAct*, we see an emergent bounded rationality defined by uncertainty over outcomes. As the future becomes more uncertain, an emotional system automatically and softly kicks in to take up the slack. The subsequent diminishment of uncertainty is transmitted socially, shared between agents in a group.

The way in which connotative and denotative reasoning and action selection are trade-offs in *BayesAct* is a reflection of a Bayesian view of the mind as an active inference engine [14]. Such a viewpoint treats the mind as operating to improve the match between internal model and external environment. Improving this match is equivalent to decreasing the total number of configurations modeled, also known as the variational free energy. At the limit, the variational free energy is the same as the true free energy. Agents with better matching models avoid surprise and are better survivors. *BayesAct* proposes the connotative space as performing a (variational) approximation to the denotative space. This approximation is necessary because of the impossibility of finding a good match at the denotative level alone, and more so in social environments which are inherently harder to predict (are less valid [12]).

Related views of emotion include the identification of negative valence with increased uncertainty [12] or with change in free energy [22], or expected free energy [18]. One component of identity (esteem) is used to modulate a reward function in [28] in order to make cooperation the more salient policy in a social dilemma. These approaches show how valence (and arousal in [18]) may be related to uncertainty (more precisely to the precision of policies). *BayesAct* shows how this relationship can be linked to social psychological theory, providing a bridge to sociological analytics.

Interestingly, the tradeoffs between connotative and denotative meanings in reasoning link social psychological theorizing across many authors. The idea traces back at least to Durkheim's instrumental vs. organic solidarity [10], is also reflected in Lawler's instrumental vs. relational commitments [25] and in Bales' forward-backward dimension [3]. When denotative reasoning takes over, individualistic groups in mechanical solidarity use instrumental commitments (more rational), and will require authority to control them and force them to obey social norms (through e.g. penalties and enforcers). They are thus following "normative" commitments [25], and must be more accepting of authority (Bales' forward dimension [3]). In more diverse groups, social complexity pushes connotative reasoning to take over, and more collectivity develops in which organic solidarity and relational commitments are more salient, and groups are less accepting of authority (more of Bales' backward dimension). Such groups will self-regulate, but allow diversity in a population. Bales' indicates a correlation between forward and more conservative political beliefs, and between backward and more liberal political beliefs [3]. *Thus, from a computational sociological point of view, we are led to the suggestion that conservatives overfit, while liberals underfit.*

V. Ongoing Practical Projects

We apply our theoretical constructs in two primary application areas. First, in studying open-source development platforms such as GitHub in the context of the THEMIS.COG project (`themis-cog.ca`) [20], and second, in the development of collaborative networks for the design of technologies to assist older adults in the context of the EMOTEC project [32]. Here we focus on the first application.

Github (`github.com`) is an online platform that is primarily used for Open Source Software (OSS) development. However, GitHub is rapidly becoming the platform of choice for general-purpose collaborative efforts. GitHub contributors can be seen as forming a large social network that is loosely bound by some developers spanning multiple projects. GitHub hosts 35 million projects and 14 million collaborators, and has seen a super linear growth over the years. At first glance, GitHub appears to be a meritocracy: contributions are made by coders with varying skill levels, and projects are advanced by individual contributions according to their quality and integrity. However, on closer inspection, it appears there are many relational factors at play, and social structures that develop within and across projects have a significant impact on the progression and biases integrated into the projects [35]. A group may include a powerful member who bullies weaker members, leading to exclusions, some of which are based on factors such as race or gender. Social status within a collaborative group can play an important role in determining the direction a project takes, and hence the final software and products being used by the general public. *BayesAct* can be used to model these interactions between GitHub contributors, and to create artificial group members whose roles are to promote and enhance inclusive collaboration. Contributors are each modeled with a *BayesAct*-based agent. Comments and interactions are then analyzed for sentiment (affect/emotion) and used to learn the affective meanings and identities for each group member. These learned identities are then used to generate information about group coherence, and to make suggestions for collaborative enhancements such as the admittance of new group members, the promotion of existing group members, or the focus of attention on specific contributions. Artificial agents, also with a *BayesAct* backend, can become group members themselves, fulfilling certain roles that fill important gaps in the social order created by the group. Artificial agents with an understanding of the relational forces at play can therefore be important moderators helping to promote inclusive and efficient development [20].

Emotionally aware agents can be useful across a wide range of other application areas, including mechanism design, behavioral economics, games, and conversational agents. Our

aim is to design and build a framework based on social-psychological theory that allows such agents to be constructed and deployed across these application areas.

VI. Ethical Considerations

The moral machine experiment [2] showed that people have shared behaviours as moral decision makers, with consistency across, and diversity within, a culture. We present a possible model for this in *BayesAct*, with this consistency arising in a sentiment (connotative) space with a simple prior distribution. This connotative space and associated temporal dynamics has a direct multimodal (emotional) communication channel providing it with information, and is learned through interaction with a social group. With a connotative space and dynamics which are consistent with others', agents can benefit by having easier focus on some aspects of the denotative world, specifically those that are relevant as solutions to social dilemmas. When they are able to follow these prescriptions for dilemmas, they become "members" of the social group in which they are learning. Thus, any moral decisions made by the agent would be consistent with those made in its social group, and therefore be acceptable. Inconsistent agent behaviours result in the ostracism of offending agents, communicated with emotional signaling and less cooperative behaviour.

VII. Conclusion

In this paper, we have proposed *BayesAct* as a computational dual-process model of human group interactions, and shown how it explicitly represents a tradeoff between the uncertainty in the denotative space (of e.g. symbolic constructs about the physics of the world) and in the connotative space (of e.g. feelings about identities and behaviours). We have argued that *BayesAct* captures some of the key elements of known human dual-process reasoning, and argued that it can be used to build artificial agents that are well aligned members of a socio-technical system.

References

[1] Jens Ambrasat, Christian von Scheve, Gesche Schauenburg, Markus Conrad, and Tobias Schrder. Unpacking the habitus: Meaning making across lifestyles. *Sociological Forum*, 31(4):994–1017, 2016.

[2] Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-Franois Bonnefon, and Iyad Rahwan. The moral machine experiment. *Nature*, 563:59–64, october 2018.

[3] Robert Freed Bales. *Social Interaction Systems: Theory and Measurement*. Transaction Publishers, New Brunswick, NJ, 1999.

[4] Lisa Feldman Barrett. The theory of constructed emotion: An active inference account of interoception and categorization. *Social Cognitive and Affective Neuroscience*, 12(1):1–23, 2017.

[5] Pierre Bourdieu. *The Logic of Practice*. Stanford University Press, 1990.

[6] Valerie Capraro and David G. Rand. Do the right thing: Preferences for moral behavior, rather than equity or efficiency per se, drive human prosociality. SSRN, May 2017.

[7] Shelly Chaiken and Yaacov Trope. *Dual-Process Theories in Social Psychology*. Guildford, New York, 1999.

[8] Francis Crick. Function of the thalamic reticular complex: the searchlight hypothesis. *Proceedings of the National Academy of Sciences*, 81:4568–4590, 1984.

[9] R.I.M. Dunbar. Neocortex size as a constraint on group size in primates. *Journal of Human Evolution*, 22(6):469 – 493, 1992.

[10] Emile Durkheim. *The Division of Labor in Society*. Free Press, 2014 (1893).

[11] Ernst Fehr and Klaus M. Schmidt. A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics*, 114(3):817–868, 1999.

[12] Oriel FeldmanHall and Amitai Shenhav. Resolving uncertainty in a social world. *Nature Human Behaviour*, In Press, 2019.

[13] Robert Freeland and Jesse Hoey. The structure of deference: Modeling occupational status using affect control theory. *American Sociological Review*, 83(2), April 2018.

[14] Karl Friston. The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11((2)):127–138, 2010.

[15] Karl Friston, S. Samothrakis, and R. Montague. Active inference and agency: optimal control without cost functions. *Biological Cybernetics*, 106(8-9):523–541, 2012.

[16] James J. Gross. The emerging field of emotion regulation: An integrative review. *Review of General Psychology*, 2(3):271–299, 1998.

[17] David R. Heise. *Expressive Order: Confirming Sentiments in Social Actions*. Springer, 2007.

[18] Casper Hesp. Hedging your bets: and active inference formulation of valence and arousal. Unpublished reserach project, 2018.

[19] Jesse Hoey, Tobias Schröder, and Areej Alhothali. Affect control processes: Intelligent affective interaction using a partially observable Markov decision process. *Artificial Intelligence*, 230:134–172, January 2016.

[20] Jesse Hoey, Tobias Schröder, Jonathan Morgan, Kimberly B Rogers, Deepak Rishi, and Meiyappan Nagappan. Artificial intelligence and social simulation: Studying group dynamics on a massive scale. *Small Group Research*, 49(6):647–683, 2018.

[21] Douglas Hofstadter. Dilemmas for superrational thinkers, leading up to a luring lottery. *Scientific American*, 248(6), June 1983.

[22] Mateus Joffily and Giorgio Coricelli. Emotional valence and the free-energy principle. *PLoS Computational Biology*, 9(6):e1003094, 2013.

[23] Daniel Kahneman and Gary Klein. Conditions for intuitive expertise: A failure to disagree. *American Psychologist*, 64(6):515–526, September 2009.

[24] Peter Kollock. Social dilemmas: the anatomy of cooperation. *Annual Review of Sociology*, 24:183–214, 1998.

[25] Edward J. Lawler, Shane R. Thye, and Jeongkoo Yoon. *Social Commitments in a Depersonalized World*. Russell Sage Foundation, 2009.

[26] N. J. MacKinnon. *Symbolic Interactionism as Affect Control*. State University of New York Press, Albany, 1994.

[27] Alexandre Mas and Enrico Moretti. Peers at work. *American Economic Review*, 99(1):112–145, 2009.

[28] Michael Moutoussis, Nelson J.Trujillo-Barreto, Wael El-Deredy, Raymond J. Dolan, and Karl J. Friston. A formal model of interpersonal inference. *Frontiers in Human Neuroscience*, 8(160), 2014.

[29] A. Ortony, D. Norman, and W. Revelle. Affect and proto-affect in effective functioning. In J. Fellous and M. Arbib, editors, *Who needs emotions: The brain meets the machine*, pages 173–202. Oxford University Press, 2005.

[30] Charles E. Osgood. On the whys and wherefores of epa. *Journal of Personality and Social Psychology*, 12:194–199, 1969.

[31] Matthew Rabin. A theory of fairness, competition and cooperation. *The American Economic Review*, 83(5):1281–1302, 1993.

[32] Julie M. Robillard and Jesse Hoey. Emotion and motivation in cognitive assistive technologies for dementia. *Computer*, 51(3), March 2018.

[33] Tobias Schröder, Jesse Hoey, and Kimberly B. Rogers. Modeling dynamic identities and uncertainty in social interactions: Bayesian affect control theory. *American Sociological Review*, 81(4), 2016.

[34] Herbert A. Simon. Motivational and emotional controls of cognition. *Psychological Review*, 74:29–39, 1967.

[35] J. Tsay, L. Dabbish, and J. Herbsleb. Influence of social and technical factors for evaluating contribution in github. In *Proc. 36th International Conference on Software Engineering*, pages 356–366, 2014.

[36] Johnathan H. Turner. The evolutionary biology and sociology of social order. In Edward J. Lawler, Shane R. Thye, and Jeongkoo Yoon, editors, *Order on the Edge of Chaos*, chapter 2, pages 18–42. Cambridge University Press, 2016.

[37] Marc Lluís Vives and Oriel FeldmanHall. Tolerance to ambiguous uncertainty predicts prosocial behaviour. *Nature communications*, 9(2156), 2018.

[38] Jing Zhu and Paul Thagard. Emotion and action. *Philosophical Psychology*, 15(1):19–36, 2002.