Judgment and Decision Making, Vol. 16, No. 2, March 2021, pp. 505-550

Denotative and connotative management of uncertainty: A computational dual-process model

Jesse Hoey* Neil J. MacKinnon† Tobias Schröder‡

Abstract

The interplay between intuitive and deliberative processing is known to be important for human decision making. As independent modes, intuitive processes can take on many forms from associative to constructive, while deliberative processes often rely on some notion of decision theoretic rationality or pattern matching. Dual process models attempt to unify these two modes based on parallel constraint networks or on socially or emotionally oriented adjustments to utility functions. This paper presents a new kind of dual process model that unifies decision theoretic deliberative reasoning with intuitive reasoning based on shared cultural affective meanings in a single Bayesian sequential model. Agents constructed according to this unified model are motivated by a combination of affective alignment (intuitive) and decision theoretic reasoning (deliberative), trading the two off as a function of the uncertainty or unpredictability of the situation. The model also provides a theoretical bridge between decision-making research and sociological symbolic interactionism. Starting with a high-level view of existing models, we advance Bayesian Affect Control Theory (*BayesACT*) as a promising new type of dual process model that explicitly and optimally (in the Bayesian sense) trades off motivation, action, beliefs and utility. We demonstrate a key component of the model as being sufficient to account for some aspects of classic cognitive biases about fairness and dissonance, and outline how this new theory relates to parallel constraint satisfaction models.

Keywords: dual-process model, emotion, affect, cognitive dissonance, fairness

*Cheriton School of Computer Science, University of Waterloo. https://orcid.org/0000-0001-5340-2204. Email: jhoey@cs.uwaterloo.ca.

†University of Guelph. https://orcid.org/0000-0002-3535-8606. Email: nmackinn@uoguelph.ca.

‡Potsdam University of Applied Sciences. https://orcid.org/0000-0002-7113-7464. Email: mail@tschroeder.eu.

We thank the editor, Andreas Glöckner, and reviewers for their comprehensive feedback on earlier drafts of the paper. This work was funded in part by the Canadian Natural Sciences and Engineering Research Council and Social Sciences and Humanities Research Council, and by the Deutsche Forschungsgemeinschaft (DFG) in Germany.

https://doi.org/10.1017/S1930297500008652 Published online by Cambridge University Press

1 Introduction

Evidence from decades of research has falsified the rational-actor view as a descriptive model of human decision-making. The limited cognitive capacity of the brain requires its frequent reliance on heuristics, emotions, intuitions and other computational shortcuts, yielding more or less adaptive choices depending on situational circumstances.[1] The present paper deals with the significance of shared cultural meanings as a constraint on decision-making, whether more or less cognitively elaborate. We know that culturally shared patterns of sense-making and narrative do influence choices (e.g., Vaisey & Valentino, 2018; Bruch & Feinberg, 2017; Kahan, 2008; Shiller, 2017) but the inclusion of such factors in formal models of decision-making has remained elusive, not least due to the enormous methodological gaps between mostly qualitative, hermeneutic studies of cultural meaning and quantitative, nomological models of decision-making.

Building on the social-psychological Affect Control Theory (Heise, 2007; MacKinnon, 1994) and our previous development of Bayesian Affect Control Theory (*BayesACT*) (Hoey et al., 2016; Schröder et al., 2016), a semantics-based model of behavior choice in uncertain social situations, we advance here the *somatic transform*, a probabilistic formalization of the relationship between denotative (well-defined, deliberate, cognitive) and connotative (vague, intuitive, affective) meanings (MacKinnon & Hoey, 2021). This new, extended, *BayesACT* model has deliberative and intuitive layers in a hierarchical (deep) network. It can therefore model deliberative heuristics and biases[2] but can also model pre-attentive, affective (emotional) biases or intuitions that are based in shared cultural beliefs, and that provide "hot" cognitions linked inextricably to synchronous "cold" cognitions. Uncertainty in the domain being modeled is precisely the element that controls the tradeoff between these two types of reasoning. *BayesACT* thus removes the requirement for any arbitrary strategy selection mechanisms, which are a problem for many decision-making models (Glöckner & Betsch, 2008).

As we will argue, *BayesACT* is broadly compatible with parallel constraint satisfaction (PCS) models (Glöckner & Betsch, 2008; Thagard, 2006), which attempt to integrate more deliberate and rational forms of decision-making with various forms of heuristic and intuitive information processing but tend to treat decision-makers as individual and disconnected from their socio-cultural environment. We argue throughout the paper that *BayesACT* is formally specified (as a Bayesian decision network), falsifiable (empirically based semantic input to the model yields precise behavioral predictions), and is more universal and at least as precise as other dual-process models (Glöckner & Betsch, 2011). In summary, *BayesACT* provides a coherent and formalized set of mechanisms for human decision-making embedded in a socio-cultural order.

[1] see https://medium.com/better-humans/cognitive-bias-cheat-sheet-55a472476b18 for a cheat sheet of over 170 documented and named cognitive biases

[2] Essentially anything that can be encoded in a decision theoretic policy of action.

https://doi.org/10.1017/S1930297500008652 Published online by Cambridge University Press

In the remainder of this paper, we first discuss the social psychological background material regarding the link between denotative and connotative meanings as reflecting cultural embeddedness of decision-makers, followed by a brief introduction to *BayesACT* and its empirical content, formality and falsifiability. We then present the mathematical details of the model, focussing on the novel operationalization of the link between intuitive and deliberative reasoning through the somatic transform. We then review other dual-process theories in Section 4, and discuss how they are related to *BayesACT*. This section includes a comparison of the PCS network models with the present *BayesACT* approach. Next, we show how *BayesACT* can be used to explain certain key facets of three foundational cognitive bias experiments, showing how this single idea can generalize across different domains: behavioral effects in cognitive dissonance, judicial decision-making, and fairness. We then conclude and sketch future work in Section 6. Technical details, code and data for *BayesACT*, including code and data for replicating the results in this paper can be found at http://bayesact.ca.

2 Theoretical Background

BayesACT is a potential mechanism for combining sociological theorising with decision making research, as it provides a parsimonious interpretation of a link between them. Connections between sociology and judgment and decision making research have been explored in Bruch & Feinberg (2017), in which a wide gulf is exposed between considering decision making as an individual process and as a social process. *BayesACT* steps in to fill this gulf, proposing a coherent link between social decision making embedded in culturally shared emotional (connotative) meanings with rational choices embedded in declarative statements about the context (denotative meanings).

While many scientific fields have approached this problem, those in economics and decision making have been amongst the most influential. Granovetter (1985) argues that economic thought is separated into two camps. First, those who consider economic action in an *oversocialized* model largely fail to predict individual action and in particular malfeasance and lying. As these theories are largely based on a relational equilibrium, they may be successful in predicting the average behavior of a group, but will not carry over to more atomistic behaviors of individuals. Second, those who consider economic action in an *undersocialized* model make predictions based on decision theoretic reasoning, but require enforceable normative rules and fail to account for relational economic behavior based on trust. Granovetter proposes an intermediate approach based on embeddedness to bridge these two camps. The social network(s) in which an agent is embedded define relational behaviors that are non-enforceable but are motivational.

BayesACT explicitly integrates a well developed *oversocialized* model of behavior, affect control theory (Heise, 2007), with a well developed *undersocialized* model based in Bayesian decision theory (von Neumann & Morgenstern, 1953). Further, it integrates

https://doi.org/10.1017/S1930297500008652 Published online by Cambridge University Press

these two theories in a principled Bayesian way, based on relative uncertainties in the two types of model. That is, the precision at which posterior beliefs can be computed decision theoretically will govern exactly how much importance this type of reasoning will have in a final posterior belief after integration with affect control theoretic reasoning. Put differently, the *relative* precision of decision theory (DT) and affect control theory (ACT) determines the relative importance of the two modes of reasoning. These precisions are dependent upon the social network in which an agent is embedded. For example, more heterogeneous groups with loosely connected nodes will tend to have less precise (more uncertain) cultural (connotative) meanings, but may still maintain precise denotative meanings due to organizational structures or task protocols imposed on the group.

2.1 Denotative and Connotative Meaning

According to the social psychological framework of symbolic interactionism, people base their decisions and actions on culturally shared meanings of things, which have evolved and are maintained and constantly reproduced in social interaction (Berger & Luckmann, 1967; Blumer, 1969; Mead, 1934). When socialized in a given culture, people internalize semantic structures that serve as default frames to keep their individual decisions more or less in line with social expectations. A socially embedded formal model of decision making needs to operationalize such meaning structures, which is precisely what *BayesACT* does. Influenced by debates in philosophy and linguistics about the distinction between *denotative* (precise, propositional, definitional) and *connotative* (vague, associative, intuitive) meanings of words for objects and events, Osgood et al. (1957, 1975) developed the semantic differential as a method for quantifying so-called "affective meanings", which were later shown to be highly consensual within linguistic communities (Ambrasat et al., 2014; Heise, 2010). These affective meanings serve in our *BayesACT* model of decision-making as a basis for a deep, intuitive layer that aligns individual decision makers with their cultural environment, while also allowing a more detached decision theoretic mode, thus avoiding the oversocialization problem.

The distinction between a connotative, socialized and a denotative, individualized process of decision making calls to memory the cognition-versus-affect debate because dual process models based on that dichotomy raise the question of the relation between these two modes of psychological experience. This is a perennial issue in psychology that reaches back almost a hundred years to the James-Lange (Lange & James, 1922) proposal that emotional experience is simply our (cognitive) perception of physiological arousal that has already occurred in response to external stimuli (e.g. we feel scared because we run, and feel happy because we smile). After a long period of relative dormancy, the issue re-surfaced in the 1980s as the primacy of cognition-versus-affect debate between Zajonc (1980, 1984) and Lazarus (1984); and continues to be a controversial subject in contemporary psychology, as evidenced by special issues of journals devoted to the topic (e.g., Mather & Fanselow, 2018).

https://doi.org/10.1017/S1930297500008652 Published online by Cambridge University Press

Although evidence from neuroscience does not support a clear distinction between cognition and affect at the neurobiological level of the brain, there is a general consensus that the distinction holds at the psychological (experiential, phenomenological) level of the mind, e.g., (Duncan & Barrett, 2007; Barrett & Satpute, 2013). At this level, the relation between cognition and affect can be expressed by two principles (MacKinnon, 1994). According to the *principle of inextricability*, cognition and affect are overlapping constituents or processes of the mind but analytically and empirically distinct as well. To expound, the relation between cognition and affect can be best viewed as a continuum between the extremes of "cold" cognitions where the intensity of affective arousal is low and "hot" cognitions where arousal is quite pronounced; or, alternatively, at the extremes of affective experience largely unmediated by cognitive processing and that involving a high level of cognitive appraisal and reaction. According to the *principle of complementarity*, to the extent that cognition and affect are at least partially independent systems, both are required for an adequate understanding of the human mind. While the principle of inextricability is an ontological statement about the reality of the mind as we understand it, the principle of complementarity is an epistemological implication of this understanding. Finally, we believe that the debate over the relative primacy of cognition and affect largely dissolves if their relation is viewed as a reciprocal process, as suggested by many authors (e.g., Mook, 1987; Forgas, 2008). An interdisciplinary overview of the relationships between these different strands of research can be found in MacKinnon & Hoey (2021).

2.2 Affect Control Theory

The reciprocal relationship between cognition and affect is a core assumption of Affect Control Theory (ACT) (Heise, 2007; MacKinnon, 1994), which capitalizes on the three-dimensional model of affective meaning established by Osgood et al. (1957, 1975). Comprising feelings of evaluation, potency, and activity (EPA), the dimensional simplicity of connotative-affective meaning provides a portal into the dimensionally complex denotative-cognitive representations of the world. Affective reactions to external objects and stimuli become "the means by which information about the external world is translated into an internal code or representation that can be used to safely navigate the world" (Duncan & Barrett, 2007, p. 1186). For example, a person wearing a lab coat in a hospital setting would lead to a denotative impression 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 are usually associated with feelings of positive valence and positive power. EPA space has been found through decades of research to be a cross-culturally normative representation of connotative meaning (Osgood, 1969).

Population surveys on semantic differential scales (with opposing adjectives at each end) provide a link from denotative to connotative as a set of samples from a distribution in sentiment space. A parametric representation can be computed (e.g. as the mean and variance

https://doi.org/10.1017/S1930297500008652 Published online by Cambridge University Press

of a normal distribution), or a non-parametric representation used (e.g. a set of samples). Parametric representations have been compiled in "dictionaries" of mappings from labels to sentiment. These dictionaries typically only code the mean of a distribution, and filtering operations are used to remove data that is not distributed normally. Thus, in ACT, a *doctor* is represented connotatively as (EPA:$\{2.7, 3.0, 0.23\}$).[3] A denotative label can be assigned to a connotative (EPA) vector in ACT using a simple nearest neighbour method (e.g. the closest label to (EPA:$\{1.5, -0.50, -2.0\}$) is *librarian* (EPA:$\{2.0, -0.46, -2.1\}$) - at a Euclidean distance of 0.017).

These surveys also yield equations for in-context impression formation. For example, participants are asked about specific actor-behavior-object (ABO) situations, e.g., *librarian reprimand* (EPA:$\{-0.36, 1.7, 1.0\}$) *bookworm* (EPA:$\{1.6, 0.41, -2.4\}$), and asked to rate each element. The basic premise of ACT is that such situations are assessed as a single unit in the connotative space, called a "transient impression". The difference between this transient impression and the out-of-context estimates (fundamental sentiments), used as an optimization loss, guides agents' actions or reinterpretations of the situation to reduce emotional incoherence. This emotional incoherence is called "deflection" in ACT. In *BayesACT*, there is additional incoherence in the denotative state, termed *ambiguity* (see Section 2.3).

The difference between fundamental sentiments and transient impressions is modeled using a set of polynomial features that multiply together aspects of the situation to yield a final estimate of the transient meanings. For example, a positively weighted polynomial term in the equation for the transient impression of an actor's evaluation multiplies the actor's fundamental evaluation score with the selected behavior's evaluation, and represents the fact that good people normally do good things. Other factors represent balance effects (good actors can do bad things to bad actors), and other elements of emotional coherence. The number of factors is empirically determined (Heise, 2010).

In the bookworm example, the *librarian*, having performed a negative and powerful behavior, results in him feeling more bad and more powerful (EPA:$\{-1.0, 1.9, 3.1\}$) with closest label *hot shot* (EPA:$\{-1.0, 1.1, 2.5\}$), while the *bookworm* would feel more bad and more active (EPA:$\{-1.2, 0.30, -0.10\}$) with closest label *truant* (EPA:$\{-0.73, -0.08, 0.40\}$).[4] Each agent can either re-identify the other or the self (as *truant* and *hot shot*), or can take action to resolve incoherence. The optimal action in this case (in the sense of reducing

[3]In the following, we use words in *italic* font for all denotative labels (symbols) that are empirically measured. For historical reasons, EPA measurements lie between -4.3 and +4.3. Unless otherwise noted, all data in this paper is taken from a survey of 1742 people in the USA, conducted at the University of Georgia in 2015. We also use data from dataset collected in 2005 at Indiana University where indicated. See https://research.franklin.uga.edu/act/ or http://bayesact.ca for access to all datasets currently available.

[4]When finding the closest neighbour in EPA space, we report the closest institutionally correct label (Heise, 2007). In this example, the closest label to (EPA:$\{-1.2, 0.30, -0.10\}$) is, in fact, *divorcée* (at a distance of 0.07), but would be denotatively implausible given the institution (a library presumably). The next closest are *communist* (distance 0.24) and *ex-girlfriend* (distance 0.38), also institutionally implausible. *Truant* is next at distance 0.43. The annotation of labels with institutions is another aspect of the model that connects connotative and denotative elements, but we do not consider it further here.

https://doi.org/10.1017/S1930297500008652 Published online by Cambridge University Press

incoherence) is to *seek advice from* (EPA:$\{2.1, 1.1, -0.34\}$) for the *librarian* and to *assist* (EPA:$\{3.3, 2.6, 0.56\}$) for the *bookworm*.

The surveys conducted under the ACT paradigm also yield the impressions of adjective-noun combinations. That is, identities can be "modified" by adjectives, and so a *doctor* who is *frustrated* (EPA:$\{-2.0, -0.34, 1.2\}$) will have a different emotional signature from a *doctor* (*frustrated doctor* EPA:$\{-1.1, 1.6, 0.70\}$). The effects of these modifiers are also determined through population surveys of modifier words. Other components of ACT may include settings (e.g. doctor in a hospital vs. doctor at home).

Emotions in ACT are defined as the vector difference between the fundamental (out-of-context) sentiments about a person and the transient (in-context) impressions formed as a result of an event. Emotions 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 causing intuitive decisions about appropriate restorative behavior. For example, if a *doctor talks down to* (EPA:$\{-1.6, -0.07, 0.31\}$) another *doctor*, the object agent is made to feel less powerful than expected (drops to -0.1), and will display *exasperation* or *indignance*. Upon receiving this signal, the acting agent may restore fundamental sentiments by *making up with* the other. Identities have "characteristic emotions" which are those felt when agents are in a perfectly coherent environment, or one that is responding exactly according to their world model. For example, a *doctor* has a characteristic emotion (EPA:$\{2.0, 0.90, 0.0\}$), with closest labels of *captivated* or *agreeable*, while for a *delinquent* it is (EPA:$\{-2.0, -0.70, 0.075\}$), with closest label *inconsiderate*.

Lastly, the mean sentiments recorded in dictionaries are empirical estimates of the distribution of the individually reported sentiment in the population studied. Thus, one would expect that measurements of these sentiments over more diverse populations would have larger variances. However, we are using these distributions as models of individual reasoning in this paper. This is reasonable given the assumption that the human mind is a model of the system it is embedded in (the "good regulator" theorem (Conant & Ashby, 1970)), and thus represents the diversity in the population as a distribution in its own model with a matching variance. Another way to see this is by noticing that a person interacting with its world will "sample" the sentiments of the network(s) in which it is embedded, and estimate the mean and variance of the sentiment in much the same way as we estimate these numbers from population surveys. Agents embedded in novel networks will potentially need to adjust their estimates in order to better "fit" the group.

2.3 BayesACT

Extending ACT, *BayesACT* is a computational dual process model of intelligence (Hoey et al., 2016; Schröder et al., 2016). One process (connotative) is continuous in a space spanning one or more dimensions and is equated with sentiment or affect, while the other (denotative) is defined on a discrete and high dimensional space and models logical and

https://doi.org/10.1017/S1930297500008652 Published online by Cambridge University Press

production-rule based reasoning. The connotative process is a probabilistic generalization of the emotional-coherence mechanism of ACT explained in the previous section, treating EPA meanings as probability distributions rather than point estimates and thus allowing uncertainty or individual variability. The dual process is built to handle uncertainty and surprise. It naturally shifts between higher bias models (lower variance in the connotative space) in more denotatively uncertain (invalid/unpredictable/surprising) situations, to lower bias models (higher precision in the denotative space) in more denotatively certain (valid/predictable/unsurprising) environments.

2.3.1 Overview

The denotative component in *BayesACT* is formulated using a probabilistic graphical model that instantiates a *temporal frame* or *structural representation* (Russell & Norvig, 2010). Frames are a classic structure used in early artificial intelligence (AI) research that assigns a label and interpretation to each object, fact, relation and event that constitute a particular situation. Such structures are typically logical and discrete-valued to enable ease of use in a computer program. For example, we might label the positions of pieces on a chess board, or predictions about how a game will turn out given a sequence of moves, or the bids in a negotiation. Frames are the foundation of much knowledge representation work in AI, and have been extended to situations with uncertainty using Bayesian networks (BNs), which compute a distribution over all possible worlds modeled by a particular frame (Pearl, 1988). This probabilistic model then rests on the structural ontology and temporal logics that are proposed in the frame. For example, in a factory assembly task, the frame may contain all objects that must be assembled, the possible orders of assembly, and the expectations about how the assembly task will proceed given a certain policy of action (e.g. for an assembly robot). Bayesian decision networks generalize the *goals* in frames as preference functions that rank all possible outcomes using a numeric scale, e.g., a utility function (von Neumann & Morgenstern, 1953). Markov decision processes (MDPs), on which *BayesACT* is based, can be represented as Bayesian decision networks, a model extensively used in operations research (Puterman, 1994).

Heuristic "expertise", routines, or social norms are also denotative as they represent default rules that are followed by the decision maker based on an appraisal of other aspects of the denotative state. Examples are stopping at a stop sign, eating dinner at six, doing what everyone else does, or buying stock when the price is low. Connotative representations model affective meanings, and lie in the low-dimensional continuous EPA space of affect control theory. Examples are the tendency for a *patient* to do something mildly deferential such as *listen to* a *doctor*, or the feeling that someone in a "hoodie" is going to commit a crime. Denotative representations are associated with cognitive processing and deliberative reasoning; connotative representations with affective processing and detection of emotional consistency between expectations and actual occurrences. So-called "intuitive reasoning" (Glöckner & Witteman, 2009) can occur in both denotative and connotative

https://doi.org/10.1017/S1930297500008652 Published online by Cambridge University Press

systems, however, with matching and constructive intuitions being handled denotatively, while associative and accumulative intuitions are handled connotatively.

Importantly, connotative meanings extend to agent actions and provide a rough (heuristic) guide over policies (i.e. search strategies). The social intelligence provided by connotative consistency is shared by agents in a community, and motivates them to want to do things according to the same practice or "habitus", which encodes the "way we do things" (Bourdieu, 1990; Ambrasat et al., 2016). This shared practice is an approximation built to handle and alleviate the computational complexity of the social world, and guides an agent towards socially acceptable choices of behavior that can ensure more globally optimal solutions to social dilemmas. Thus, *BayesACT* implements a weighted mixture of strategies, ranging from fully connotative, socially oriented, to fully denotative, individually oriented. Emotional coherence, as in Thagard (2006), is one component of a final determination of action.

The link in *BayesACT* between denotative and connotative induces a natural (Bayesian) tradeoff due to relative uncertainty in connotative and denotative states. As the environment becomes less valid (so the distribution over denotative states is more dispersed or has higher entropy),[5] the posterior will be more heavily influenced by the prior in the connotative state. Agents in less valid (less predictable) environments will thus 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. The tradeoff also goes in the other direction simultaneously. That is, higher entropy connotative states (perhaps due to emotional signaling noise) will push reasoning towards the denotative meanings, thus having the same effect as an increase in validity in the denotative state.

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 rely more on fairness in uncertain social situations (see Section 5.1 and (van den Bos, 2001)), and act more pro-socially (cooperate in a public goods game) in ambiguous situations (ones in which risk is hard to evaluate), see Vives & FeldmanHall (2018). 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

[5]Entropy is a measure used in statistical physics, and it describes the level of homogeneity of a distribution. In information theoretic terms, entropy measures the amount of information in a system that exists in a set of states x according to the probability distribution $P(x)$. Entropy is typically written as $S(P) = -\sum_x P(x) \log P(x)$. A low entropy system has a distribution over states $P(x)$ which is not dispersed evenly across all x, and so is easier to predict (or requires fewer bits of information to transmit on an information channel). Kahneman & Klein (2009) refer to this as the "validity" of the environment, but note that this also depends on the expertise of the agent. That is, experts can have a valid model of what seems like an invalid environment to most non-experts, as they have constructed a more complex (higher capacity) model.

https://doi.org/10.1017/S1930297500008652 Published online by Cambridge University Press

more well defined, and so ambiguity (the uncertainty in risk) is lower. A major contribution of the present paper is to clarify and formally specify these concepts through introduction of the somatic transform (see Section 3).

2.3.2 Mathematical Model

BayesACT is a Bayesian decision network known as a partially observable Markov decision process (POMDP), a formal model with precise probabilistic and decision theoretic semantics (Åström, 1965; Puterman, 1994; Kaelbling et al., 1998; Boutilier et al., 1999). In the following, we use standard notation where capital letters denote random variables, small letters represent values those variables can take on, and bold letters represent sets of variables or values. Thus, $\mathbf{X}$ is a set of random variables, X is a single random variable, $\mathbf{x}$ is a particular value for all variables in the set $\mathbf{X}$ and x is a particular value that variable X can take on. Capital letters in calligraphic script, $\mathcal{X}$, denote the space spanned by a variable or set of variables.

We start with the assumption that an agent must maintain an internal model of the world as a set of states making up a state space $\mathcal{S}$, which is factored into a *denotative* part, $\mathcal{X}$, that describes the ontological states of entities in the world, and a *connotative* part, $\mathcal{Y}$, that describes the affective meanings of entities in the world. In ACT and *BayesACT*, the connotative space spans the three-dimensional vector space of EPA sentiments for identities (labels assigned to people) and behaviors (labels assigned to people's actions). The denotative space is discrete and factored, and contains representations of each actor's identity (as a label). For example, a particular agent's identity may be represented by some X, such that the word *doctor* in the English language is a particular value of that variable $X = doctor$. Similarly, $\mathbf{Y} = y$ will represent the affective meaning (in EPA space) of that denotative entity.[6]

BayesACT has two random variables corresponding to observations, $\mathbf{\Omega} = \{\mathbf{\Omega_x}, \mathbf{\Omega_e}\}$. One, $\mathbf{\Omega_x}$, represents signals about the environment giving evidence for the denotative state. The other, $\mathbf{\Omega_e}$, 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. Finally, *BayesACT* has two sets of actions representing denotative action, A, and connotative

[6]Each of the variables $\mathbf{X}$ and $\mathbf{Y}$ have multiple components. For instance, $\mathbf{Y}$ is in fact a set of 18 different random variables, one for each actor-behavior-object crossed with evaluation-potency-activity for each of the two sentiment types (fundamental sentiments and transient impressions). $\mathbf{X}$ may have as many components as required by the domain, but at least represents actor, behavior and object identities as discrete-valued variables ranging over identity labels. Institutional and other logical constraints may be placed on the dynamics of $\mathbf{X}$ depending on the problem at hand (Hoey et al., 2016). In this paper, our focus on the connection between denotative and connotative allows us to abstract this away and consider a single variable X and Y.

https://doi.org/10.1017/S1930297500008652 Published online by Cambridge University Press

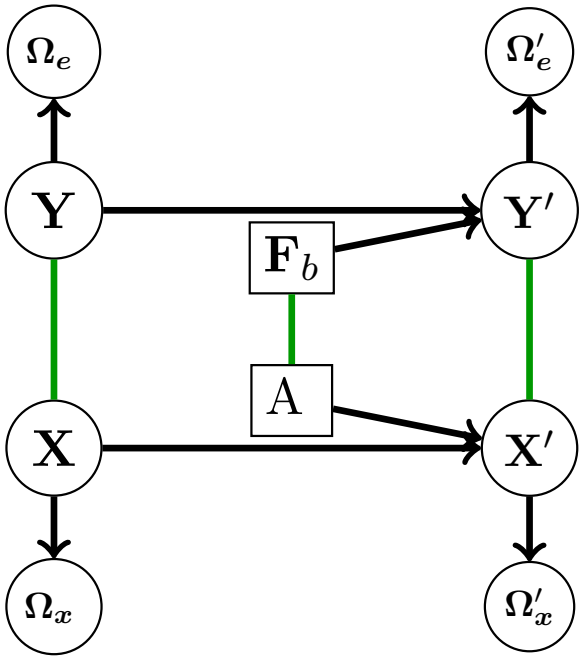


FIGURE 1: Bayesian decision network for *BayesACT* at a high level of abstraction showing denotative $\mathbf{X}$ and connotative $\mathbf{Y}$, observations $\mathbf{\Omega_x}$, emotions $\mathbf{\Omega_e}$, and actions both denotative A, and connotative $\mathbf{F}_b$. Directed links imply (but do not require) causality as in a Bayesian network. Two somatic potentials, modeled as undirected links (see text), link state and action, respectively. Primed variables are post-event, and the network is dynamically unrolled through time.

meanings of those actions, $\mathbf{F}_b$.[7]

Figure 1 shows a graphical model representation of *BayesACT*.[8] This graphical model shows random variables as circles, decision variables (under the control of the agent) as squares.[9] Undirected (directed) links give probabilistic (conditional) dependencies between variables. Thus, an undirected link between X and Y represents $P(X,Y)$, while $X \rightarrow Y$ represents $P(Y|X)$.

The graphical model is defined for a discrete temporal sequence, with one set of variables for each time step. Two time slices are shown in Figure 1, with primed variables indicating those that are in existence after an action is taken, but in reality the network is "unrolled" for as many time steps as required, and the dynamics is considered stationary (does not change over time).[10] A Bayesian decision network, along with associated probabilities

[7]The connotative action is called $\mathbf{F}_b$ because it is the behavior component of the fundamental sentiment, $\mathbf{F}$. See Hoey et al. (2016) for details.

[8]Such models are a formal representation which is useful for human prior specification of the domain, the standard reference is Pearl (1988), although treatments can be found in any artificial intelligence textbook (Russell & Norvig, 2010), the most comprehensive being Koller & Friedman (2009). For MDP-specific material, see Puterman (1994). For POMDP-specific material refer to Boutilier et al. (1999).

[9]Rewards or utility functions are usually shown as diamonds, but are not shown in Figure 1 to reduce clutter.

[10]Although by deepening the network, one can account for changes in, and learning of, these dynamics,

https://doi.org/10.1017/S1930297500008652 Published online by Cambridge University Press

and utility functions, can be used to answer queries of the form $P(\mathbf{X}_t|\mathbf{\Omega}_0, \mathbf{\Omega}_1, \ldots, \mathbf{\Omega}_t)$, the posterior distribution over the denotative state at time t given observations up to that time of $\mathbf{\Omega}_0, \mathbf{\Omega}_1, \ldots, \mathbf{\Omega}_t$. A number of algorithms exist for computing these inferences, and these can be computed efficiently and recursively for structured models such as a POMDP by repeatedly taking products and sums of conditional probability and utility functions. Optimal decisions can also be computed using similar algorithms, where maximization over the action variables takes place of summations for random variables. However, these computations are significantly more difficult than those for inference, and normally require approximations such as stochastic simulation (Monte-Carlo tree search), or variational inference, see (Hoey et al., 2016; Asghar & Hoey, 2015; Silver & Veness, 2010).

BayesACT is formally specified (as a dynamic decision network) and makes clear predictions about tradeoffs between making decisions based on affective identity and expected utility. It therefore has a clear specification and empirical content (Glöckner & Betsch, 2011). Further, it is falsifiable as identity can be modified as an independent variable to make different predictions with different experimental setups (see Section 5.3). In high certainty conditions, we expect decisions to be largely independent of identity, so *BayesACT* predictions would be largely in agreement with parallel constraint network models. In low certainty conditions (high ambiguity), we expect the predictions of the two models to diverge, but we expect the *BayesACT* predictions to be more precise due to its ability to specialize based on this additional context.

3 The Somatic Transform: Modeling Denotative and Connotative Management of Uncertainty

In the original formulation of the *BayesACT* model, the sentiment of the behavior being performed was directly observed in an interaction as a three-dimensional, continuous vector. That is, if a doctor was observed injecting someone with medicine, then *BayesACT* and ACT both expected a direct observation of the mean EPA rating for that denotative behavior, *inject someone with medicine*: (EPA:$\{0.91, 1.7, -0.23\}$). *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. In the medical domain, this may include the weighing of the relative risks and harms of a prescription drug based on available scientific evidence, for example.

Here, we propose that a *BayesACT* agent also includes a denotative representation of identities and behaviors of other agents. Rather than a single label applied to a situation (e.g. *doctor*), *BayesACT* maintains multiple possible labels (e.g. *doctor*, *nurse*) and also maintains a probability distribution over them. In this case, these denotative elements are

leading to a hierarchical POMDP that can be treated as a Bayesian non-parametric model - the infinite POMDP (Doshi-Velez, 2009) - and Bayesian reinforcement learning (Duff, 2002).

https://doi.org/10.1017/S1930297500008652 Published online by Cambridge University Press

linked to the connotative state through a potential function that measures the incoherence (difference) between the posterior estimate of the denotative state (e.g. *doctor*) and the posterior estimate of the connotative state (a distribution in the affective EPA space). We call this potential function the *somatic potential*, defined as an energy-like measure of the difference (incoherence) between the connotative and denotative spaces.

For example, if the doctor performs some behavior uncharacteristic of a doctor (e.g., *harass* 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 behavior pushes the observing agent to a higher energy state. While behaviors 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. For example, if a *doctor* is observed *harassing* (EPA:$\{-3.0, 0.55, 1.6\}$) a *patient* (EPA:$\{0.64, -1.5, -1.3\}$), agents would be motivated to act in such a way as to stop the behavior, 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 Euclidean distance of 0.44).

In the following, we present a mathematical definition of the somatic potential and associated energy function. Recall that our objective in this paper is to study the probabilistic inferences that model the trade-offs between connotative and denotative states in decision making. We therefore abstract away much of the *BayesACT* model as presented in this section (which, as noted above, is itself an abstraction of the full model). A mathematical presentation of this abstraction is provided in the Appendix, and the interested reader is referred to Hoey et al. (2016) for further details on the elements we are leaving out here.

3.1 The Somatic Transform

A core element in the *BayesACT* POMDP is that every denotative element $\mathbf{X}$ (including agent actions, A) has an associated connotative element $\mathbf{Y}$ ($\mathbf{F}_b$ for behaviours). The connection between denotative and connotative elements is written as a functional called the somatic transform: $\hat{\mathbf{G}}(\mathbf{X}, \mathbf{Y})$. The somatic transform specifies the shared cultural connotative interpretation of the denotative state $\mathbf{X}$. That is, for some $\mathbf{x}$, $\hat{\mathbf{G}}(\mathbf{x}, \mathbf{Y})$ is a function over the connotative space, $\mathcal{Y}$, representing the shared sentiments for denotative state $\mathbf{x}$. For example, $\hat{\mathbf{G}}(X_a = doctor, \mathbf{Y})$ could be a normal distribution given by summary statistics (mean, covariance) measured in a population survey. Using a somatic transform, agents can interoceptively ask questions such as "what do I feel about object $\mathbf{x}$?" (e.g., Q: "how do you feel about doctors?"; A: "I see doctors as quite good, quite powerful, and a bit active, but I'm somewhat uncertain about this").

The somatic transform can also yield, for some $\mathbf{y}$, a function over the denotative space, $\mathcal{X}$, $\hat{\mathbf{G}}(\mathbf{X}, \mathbf{y})$. For some discrete set of N denotative entities, $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2 ... \mathbf{x}_N\}$, this could be a multinomial distribution (a set of numbers $p_1, p_2, ... p_N$ such that $0 \leq p_i \leq 1 \ \forall i$ and $\sum_i p_i = 1$) indicating that entity $\mathbf{x}_i$ is likely with probability p_i. Such a distribution can be

https://doi.org/10.1017/S1930297500008652 Published online by Cambridge University Press

computed using, e.g., a kernel method. The agent can therefore also ask questions such as "what sort of **x** feels like **y**?" (e.g., Q: "what sort of things are very good, a bit weak, and very active?"; A: "dogs, smiley faces, kids").

To give the somatic transform a more precise definition, we may model it as a joint probability distribution over **X** and **Y**, written as a Boltzmann distribution[11] of the somatic potential, which we write as $E(x, y)$, denoting the energy or incoherence between the denotative and connotative states:

$$G(x, y) = ce^{-E(x,y)} = c\ e^{-(y-M(x))^2/\gamma^2}, \tag{1}$$

where $M(X = x)$ is a function in Y giving the connotative value for that particular x and c is a normalizing constant (the inverse of the partition function). M may be simply the mean of the population survey for the concept $X = x$, and γ a constant. Thus, the more coherent x is with y, the smaller the energy $E(x, y) = (y - M(x))^2/\gamma^2$, and the more likely the state. The parameter γ models one aspect of the (affective) predictability of the environment. As the environment's diversity increases (say with the addition of some heterogeneous other agents), γ naturally increases as the ascription of sentiment to denotative elements in the world is less well defined. Such a world becomes less predictable and less valid both connotatively (sentiment less well defined) and denotatively (heterogeneous agents behaving in different ways). A value for γ is a learning choice to be made by an agent. Although we know that γ is related to the variance in sentiments in the population, it does not need to be exactly the same. The value of γ will also be a function of the agent's social network, as the agent may operate in a clique or cluster of locally more homogeneous sentiments.

Here, we suppose that a prior marginal over X, $P(x)$, represents the agent's current estimate of the denotative nature of the interaction. It can be a belief about various properties of objects being manipulated as part of the ongoing interaction, for example. Thus, an agent may observe a female in a white coat in a hospital setting, and infer a distribution over the possible identities of *nurse* or *doctor*. In this case, an agent with a gender stereotype may form a denotative impression which puts more mass on *nurse* than on *doctor*.[12] The result of the inference is $P(x)$, and would be represented in this simple

[11]The Boltzmann distribution is often used in statistical physics to describe the joint probability of different states (e.g. of a gas or spin system). It is usually represented as $P(x) = \frac{1}{Z}e^{-E(x)/kT}$ where $E(x)$ is the energy of configuration x, T is the temperature and Z is a normalizing factor called the "partition function". Estimating the partition function (and thus knowing the actual probability of an event x) is challenging because it is a sum of energy terms across all possible states (many of which may not even be known by the agent). The partition function is equal to the free energy of the system scaled by the temperature, T. The free energy is a measure of how complex the environment or system is. We use the Boltzmann distribution here for convenience. It could be replaced with other distributions or learned.

[12]This example is meant to be demonstrative of the primary ideas, and so starts with the assumption of such a gender stereotype, without prejudice against evidence that gender may not make stereotypes salient in technical settings (Correll et al., 2020).

https://doi.org/10.1017/S1930297500008652 Published online by Cambridge University Press

case with one marginal number p giving the probability that this entity (person) is a *nurse*, for a distribution over [*nurse*, *doctor*] of [13]

$$P(x) = P([nurse, doctor]) = [p, 1-p].$$

Suppose further that a prior marginal distribution over Y, $P(y)$, represents an agent's current estimate of the affective nature of the interaction. It can be a belief about identity sentiments of participants, recent or forthcoming behavior sentiments, or sentiments about settings or other physical objects. Suppose the agent had observed the same female in the white coat performing a gesture implying she has power, such as ordering someone to do something. In this case, the agent's prior over the sentiment about the white-coated female, Y, would be shifted towards more powerful values which might conflict with denotative bias.

Using a joint prior that is factored as $P(\mathbf{X}, \mathbf{Y}) = P(\mathbf{X})P(\mathbf{Y})$, we seek a posterior distribution $P'(\mathbf{X}, \mathbf{Y}) = P(\mathbf{X}', \mathbf{Y}')$ that combines priors with the constraint imposed by the somatic transform (Equation 1). We write both distributions and density functions as $P(\cdot)$, as the type of function and operations used is defined by the variable in context $\mathbf{X}$ or $\mathbf{Y}$. As we are focussing on the somatic transform only, we ignore the temporal dynamics (explored more fully in Hoey et al. (2016)), and so consider primed and unprimed variables to be the same (so the prior over unprimed variables becomes the prior over primed variables). We further ignore the observation spaces $\mathbf{\Omega_x}$ and $\mathbf{\Omega_e}$, simply integrating probability weights into priors over $\mathbf{X}'$ and $\mathbf{Y}'$, respectively. In fact, we are seeking the posterior *given our knowledge of Equation 1*. We postulate a variable G representing our knowledge that a somatic transform such as Equation 1 connects denotative ($\mathbf{X}$) and connotative ($\mathbf{Y}$) spaces. We therefore write that

$$P(\mathbf{X}', \mathbf{Y}') \equiv P(\mathbf{X}', \mathbf{Y}'|G) \propto \hat{\mathbf{G}}(\mathbf{X}', \mathbf{Y}')P(\mathbf{X}')P(\mathbf{Y}'), \tag{2}$$

and we see that the joint distribution is in fact a product of the somatic transform and the posterior distribution (which may be factored into two distributions over $\mathbf{X}'$ and $\mathbf{Y}'$). In the one-dimensional case, with a somatic transform defined to be the Boltzmann distribution as above, and factored priors $P(x)$ and $P(y)$, an estimate of the marginal posterior over Y ($P'(y)$, the feelings evoked after an event) can be computed as:

$$P'(y) = \sum_x P'(x, y) = \sum_x P(y)P(x)G(x, y) \propto P(y) \sum_x P(x)e^{-(y-M(x))^2/\gamma^2}, \tag{3}$$

where the $\propto$ shows the quantities are proportional (equal up to a constant multiplier, labeled c in Equation 1). Thus, the posterior distribution over Y is the expectation of the somatic transform with respect to the prior distribution over X, $P(x)$, multiplied by the prior distribution over Y.

[13]Clearly there may be more than two identities under scrutiny at a time (e.g. head nurse, orderly, medical student, etc). We proceed here with two without loss of generality.

https://doi.org/10.1017/S1930297500008652 Published online by Cambridge University Press

Similarly, the marginal posterior distribution over X, $P'(x)$, is given by

$$P'(x) = \int_y P'(x,y)dy = \int_y P(y)P(x)G(X,y)dy \propto P(x)\int_y P(y)e^{-(y-M(x))^2/\gamma^2}dy. \quad (4)$$

The posterior distribution over X is the expectation of the somatic transform with respect to the prior distribution over Y, $P(y)$, multiplied by the prior distribution over X.

Equations 3 and 4 can be analytically computed in the case of Gaussian priors, which we pursue below. In practice, the somatic transform can be problematic because it generates a posterior over $\mathbf{Y}$ which is no longer a single Gaussian, but a sum of Gaussians. Projecting this very far into the future may lead to an explosion of modes. However, modes can be combined or rejected by action selection as well, meaning that for each sum of Gaussians generated, one can be selected through action. In the doctor example, a sum of two Gaussians results after a single iteration but the act of deference performed can resolve much of this uncertainty by committing to one hypothesis or the other. Further, there are analytical methods for handling the explosion of modes, such as Bayesian moment matching (Jaini & Poupart, 2016). Finally, higher level models at the institutional or self level may help through regularization (damping) of learning (MacKinnon & Heise, 2010; Hoey & Schröder, 2015).

3.2 Relationship of *BayesACT* and ACT

In Affect Control Theory (ACT), the somatic transform has a zero temperature parameter $\gamma = 0$. Further, either $P(x)$ is a point estimate (x_o), and $P(y)$ is a constant (no prior, set to 1), or $P(y)$ is a point estimate (y_o) and $P(x)$ is a constant. Writing a point estimate as a delta function, $\delta(x - x_o)$, where $\delta(x - x_o) = 1$ if $x = x_o$ and 0 otherwise, then in the first case, Equation 3 is:

$$P'(y) = \sum_x \delta(x - x_o)G(x,y) \propto e^{-(y-M(x_o))^2/\gamma^2} = \delta(y - M(x_o)), \quad (5)$$

and in the second case, Equation 4 is:

$$P'(x) = \int_y \delta(y - y_o)G(x,y) \propto e^{-(y_o-M(x))^2/\gamma^2} = \delta(y_o - M(x)), \quad (6)$$

which simply says that x_o and y_o are related through the function M directly (e.g. M is a dictionary linking each x_o with a y_o). Technically in Equation 6 this assumes a dictionary with an entry for every possible y_o, clearly an impossibility. Finding the nearest neighbor, as described above, is one possible way to circumvent this.

As mentioned previously, the somatic transform is the key difference between our presentation of the model here and the presentation of it in the original exposition of the BayesACT model (Hoey et al., 2016; Schröder et al., 2016). In the original presentation, we assumed that behaviors were both perceived and generated in the connotative/affective-space, leaving the translation to/from denotative spaces to some other perceptual or motor

https://doi.org/10.1017/S1930297500008652 Published online by Cambridge University Press

system. The somatic transform mathematically defines this translation and integrates it directly and deeply into the model itself. Rather than perceiving a behavior as a vector in a 3D affective space, an agent perceives and cognitively interprets denotative aspects of the situation including behavior, then uses these denotative aspects as evidence in support of its connotative/affective predictions, computing the level of support using a somatic transform. Simultaneously, connotative predictions are mapped to future denotative states and actions, providing heuristic guidance to an agent. This aspect of the original presentation of the BayesACT model is thus revealed as a simplifying assumption that has been replaced here with the more general somatic transform.

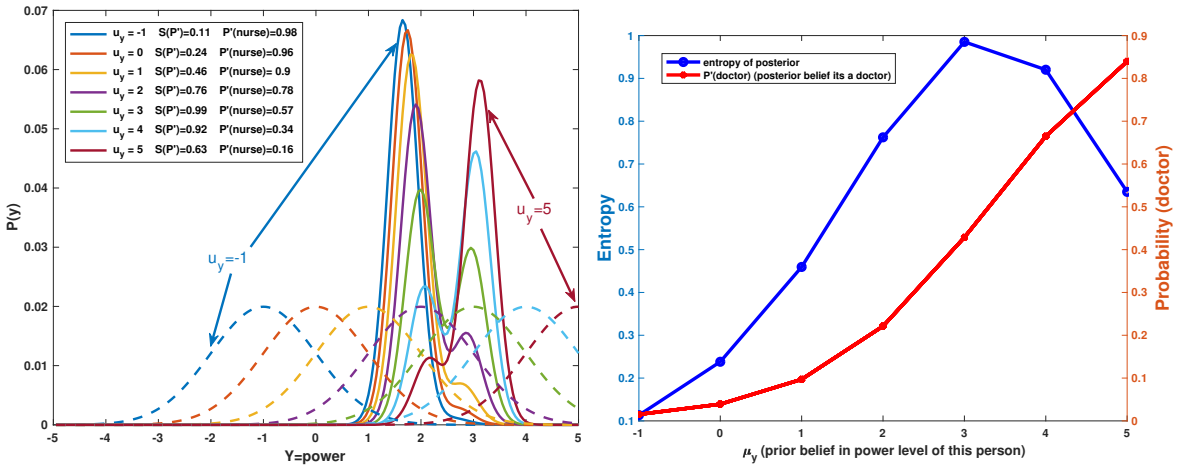
3.3 Somatic Transform Examples

Figure 2 shows an example usage of this transform for the simple case discussed previously. In it, we consider a prior over y, $P(y)$, as a Gaussian distribution with a variance $\sigma_y = 2.0$ and a mean u_y which is varied to see the effect of a changing connotative prior. These distributions are shown as dashed lines in Figure 2 (with means ranging from -1 to 5). A prior over x, $P(x)$, represents only two identities *nurse* and *doctor* with probabilities 0.7 (nurse) and 0.3 (doctor). A-priori, the agent believes it is more likely this person is a nurse. The somatic transform is implemented using normal distributions as the values of M mapping a label in X to a mean and variance in Y given by suurvey data generated at the University of Georgia in 2015. The identities of *nurse* and *doctor* have power (P) values of 1.9 and 3.0, respectively. We only consider a single dimension here for ease of exposition, but the results carry over to 3 (or, in principle, more) dimensions. In Figure 2, we use $\gamma = 0.3$, but investigate how γ affects the results in Figure 3.

Figure 2(a) shows how the posteriors evolve as μ_y is changed. The entropy in the posterior distribution over x, $P'(x)$, is shown as $S(P')$ in the legend and in Figure 2(b), along with the value of $P'(X = nurse)$ as $P'(nurse)$. The posteriors over y, $P'(y)$ are shown as solid lines. First, we can see that as the prior over y approaches the prior over x (with an expected Power sentiment value of $0.7 \times 1.9 + 0.3 \times 2.95 = 2.2$), the posterior becomes a bimodal distribution with about 70% of its mass nearer to the *nurse* identity at 1.9. Further, as the priors more closely agree (that is μ_y approaches 2.2), the entropy of the resulting distribution over x increases, so the information obtained by combining them is smaller. For largely different values of the mean of $P(y)$ (e.g. $\mu_y = -1.0$ or $\mu_y = 5.0$), the resulting entropy of $P'(x)$ is small, and more information was gained by the denotative system from the connotative system. The resulting distributions over x are also shown, demonstrating a clear shift from *nurse* ($P'(x) \rightarrow 1.0$) to *doctor* ($P'(x) \rightarrow 0.0$) as the prior information about the sentiment observed shifts towards the positive (the person demonstrates a behavior with more power). Figure 2(b) shows the denotative entropy and posterior as a function of μ_y, demonstrating the trends with more clarity.

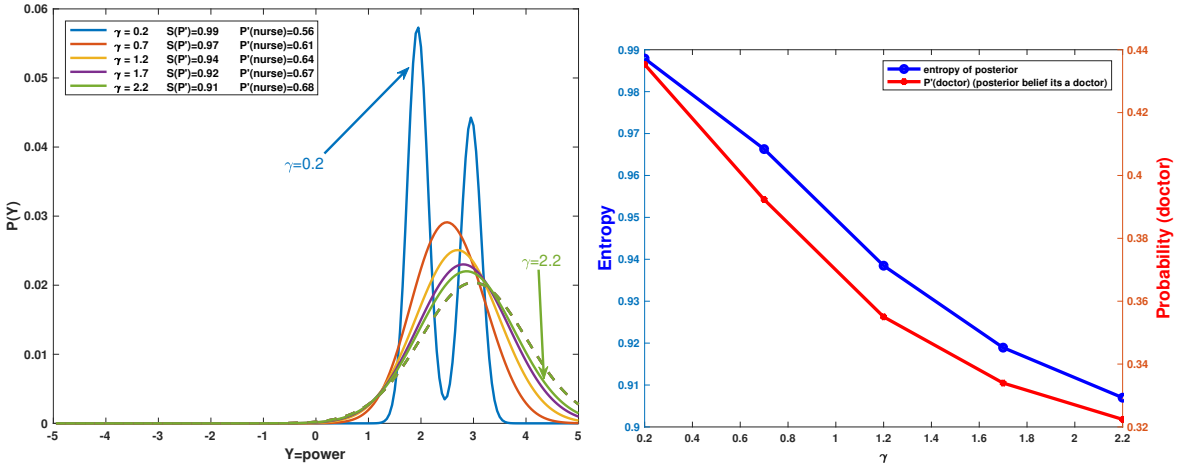
Figure 3(a) shows how the same curves evolve according to a changing value of γ, with a fixed $\mu_y = 3.0, \sigma_y = 1.0$ (dashed line), $P(X = nurse) = 0.7$. As Figure 3 shows, when γ

https://doi.org/10.1017/S1930297500008652 Published online by Cambridge University Press



(a) (b)

FIGURE 2: Effects of the somatic transform on the posterior marginals over X and Y. (a) Gaussian priors over Y are shown as dashed lines for different values of μ_y. The prior over X is $P(X = nurse) = 0.7$. The posterior over Y is shown as solid lines, while the posterior over X is shown in the legend, with $S(P')$ denoting the entropy of $P'(X)$ and $P'(nurse)$ denoting $P'(X = nurse)$. As the prior shifts to more positive values in Y, the posterior in Y shifts to be more in line with the power sentiment about *doctor*, rather than *nurse*. Further, the posterior in X also favors *doctor* (that is, $P'(nurse) \rightarrow 0.0$). (b) plot of the overall trends in entropy and posterior probability as a function of μ_y.



(a) (b)

FIGURE 3: Posterior over Y with varying γ. (a) As γ decreases, the posterior over Y is more focussed on the priors over x. $S(P')$ denotes the entropy of $P'(X)$. $P'(nurse)$ denotes the posterior probability of X being *nurse*: $P'(X = nurse)$. Prior in Y is $\mu_y = 3.0, \sigma_y = 2.0$. (b) plot of the overall trends in entropy and posterior probability as a function of γ. Note the different scale in (b) to Figure 2.

https://doi.org/10.1017/S1930297500008652 Published online by Cambridge University Press

is large (> 2.0 in this example) there is not as strong an effect between X and Y, and so both follow their prior distributions closely. For small γ, denotative and connotative are more strongly linked, so the sentiment follows the prior over X more closely, becoming more centered around the known mean values of power for the identities of *nurse* and *doctor* of 1.9 and 3.0. The denotative posterior is shifted towards $X = doctor$, as the connotative prior indicates a powerful identity.

The somatic transform naturally shows a trade-off between the uncertainty in X and Y. E.g., Figure 4 shows how the posterior over X and Y changes as a function of $P(X = nurse)$. In this simulation $\mu_y = 3.5, \sigma_y = 1.0$ and $\gamma = 0.2$. As the environment becomes less valid (less predictable or more uncertain, so $P(x)$ is more dispersed or has higher entropy), then $P'(y)$ and $P'(x)$ will be more heavily influenced by the prior in y: they will make inferences and choose actions that are more in line with connotative (socio-cultural) expectations. In more valid environments, $P(x)$ has lower entropy and dominates the posterior.

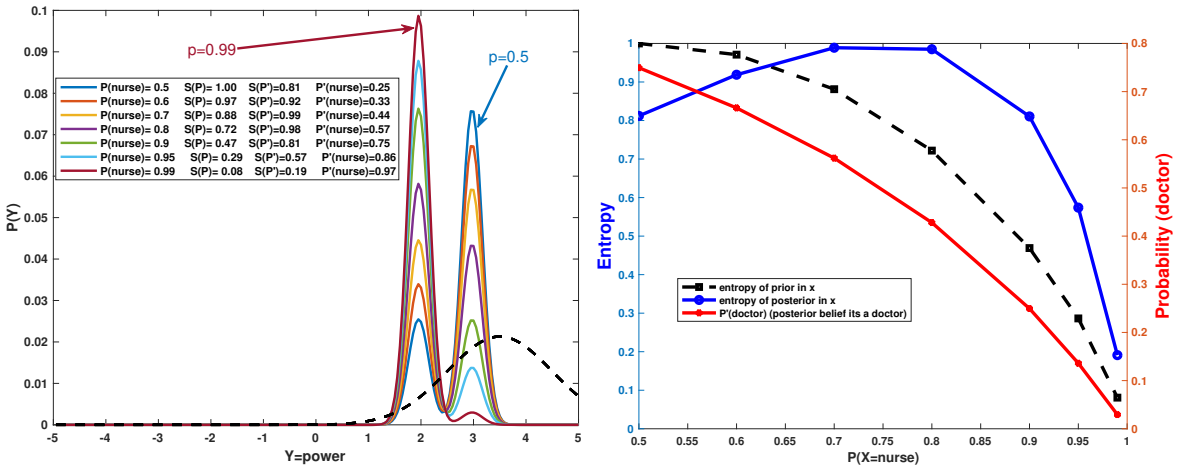
These basic elements in *BayesACT* are in line with experiments showing how humans tend to act more pro-socially (cooperate in a public goods game) in ambiguous situations (ones in which risk is hard to evaluate, see Vives & FeldmanHall (2018)). In this simulation, risk is $p = P(X = nurse)$: if $[p, 1 - p]$ is lower entropy, then risk is more well defined, and so ambiguity (the uncertainty in risk) is lower. It is also in line with experiments showing how the effects of increased ambiguity may result in more political polarization (Bail et al., 2018). As people are exposed to more heterogeneous discussion online, their political views become more deeply entrenched. More homogeneous discussion leads to more muted political beliefs being expressed. This is explainable by noting that the more heterogeneous discussion leads to greater denotative uncertainty, leading to a more heavy reliance on connotative priors, which may have stereotypes embedded into them. As discussion becomes more homogeneous, increased weight is placed on available evidence, leading to a posterior biased towards a more neutral solution (assuming the evidence is independent of the arguments presented and the stereotypes held).

4 Related Dual Process Models

In this section, we discuss the relation of the *BayesACT* model with other theories of dual (or multiple) processes. We will not provide a detailed review of such approaches, which are ample, but in many cases not formally implemented in computational models. We will instead focus on a more in-depth comparison of our model with the class of parallel constraint satisfaction (PCS) network models, which bear many conceptual similarities with *BayesACT* but also a few notable differences. The same is true for more recent biologically inspired computational dual-process models, which we will also briefly contrast with the *BayesACT* approach.

Dual process theories are well studied in social psychology (Chaiken & Trope, 1999), but many different terms are used to refer to the two levels of processing. "Cognitive" processing

https://doi.org/10.1017/S1930297500008652 Published online by Cambridge University Press



(a) (b)

FIGURE 4: (a) The red line shows a prior state with a less dispersed $P(X = nurse) = p$ with $p = 0.99$, respectively, yielding a posterior for both X and Y that is more in line with the original denotative prior $P(x)$. The blue line ($p = 0.5$) shows how the posterior is biased towards the prior in y (possibly based on stereotypes). The prior in y is shown as a black dashed line (same for all values of p). $S(P)$ and $S(P')$ denote the prior and posterior entropy of $P(X)$, and $P(nurse)$ and $P'(nurse)$ denote the prior and posterior probability of X being *nurse*. (b) plot of the overall trends in entropy and posterior probability as a function of $P(X = nurse)$.

is often referred to as deliberative, reflective (Ortony et al., 2005), conscious (Smith et al., 2019) or "System 2" (Stanovich & West, 2000), whereas "emotional" processing is called automatic, routine (Ortony et al., 2005), or "System 1" (Stanovich & West, 2000). The machine learning community has also started to discuss "System 1" and "System 2" thinking, although the interpretation of these terms is much more functional in that "System 1" is relegated to basic pattern recognition as instantiated in deep learning.[14] The ability of deep neural networks to implement symbolic-like deliberative reasoning is a popular current topic of research (Rohekar et al., 2018). Behavioral economists have brushed against computational dual-process models by proposing a variety of mechanisms that explain the experimental evidence of pro-social (e.g. cooperative) behavior in humans. Early work on motivational choice (Messick & McClintock, 1968) proposed a probabilistic relationship between game outcomes (payoffs) and cooperative behavior. This led to the proposition that humans make choices based on a modified utility function that includes some reward for fairness (Rabin, 1993) or penalty for inequity (Fehr & Schmidt, 1999). Modifications to the utility function based on *identity* have also been proposed (Akerlof & Kranton, 2000). However, these identities are denotative labels with associated situation-specific and static behaviors that are defined on a case-by-case basis, and the utility change produced

[14]See the recent NeurIPS keynote by Bengio on this topic https://www.youtube.com/watch?v=FtUbMG3rlFs

https://doi.org/10.1017/S1930297500008652 Published online by Cambridge University Press

by an identity-behavior combination is not well defined (although conceptually a crucial aspect of the model). Further, it appears that fairness or inequity adjustments may not be comprehensive enough to account for human behavior across all games, and a morality concept that is not based on outcomes can be used as a more parsimonious account (Capraro & Rand, 2018; Haidt, 2001). The question of how this morality is defined is left open, but *BayesACT* may provide such a link to sociological theorising about morality through the ACT lens.

Evans (2008) differentiates between *parallel-competitive* (PC) and *default-interventionist* (DI) dual process models. In PC, instrumental (System 2) plans of action are, if used sufficiently, hard-coded into an associative network that can later be quickly retrieved by System 1, and which then competes with ongoing System 2 reasoning for a given situation. In DI, the System 1 process sets a context in which the System 2 process can reason. Glöckner & Witteman (2009) further decompose PC and DI into different forms of intuition, relating PC to autonomous control of behavior expressed as associative or matching intuition, and DI to pre-attentive accumulative and constructive forms of intuition. *BayesACT* has components for each of the four types of intuition, with denotative reasoning being used for matching intuition (the ability to observe and reason) and for constructive intuition (the ability to plan), while connotative reasoning is used for associative intuition (the ability to feel) and for accumulative intuition (the ability to process feelings and regulate social dynamics). In *BayesACT*, the observation functions (relating $\mathbf{\Omega_x}$ to $\mathbf{X}$ and $\mathbf{\Omega_e}$ to $\mathbf{Y}$) model sensory precision, which corresponds to *matching intuition*: exemplar classification that happens according to learned patterns. The denotative dynamics $P(\mathbf{X}'|\mathbf{X}, a)$ is able to model probabilistic dependencies through time, which allows for reasoning (propagating probabilistic beliefs *backward* through time to assign probable causes, or forwards through time to compute possible futures, which may include expected habitual actions (Wood et al., 2002)), and planning (propagating beliefs *forward* through time to locate potential rewards). These are standard operations in a POMDP. The somatic transform handles associations between denotative and connotative meanings (associative intuition defined as conditioning, or learning the value of things). Lastly, evidence is incorporated into the connotative state allowing for accumulative intuition from emotional signals. Affective factors can exist in both denotative and connotative layers, but take center stage as the primary dynamical process in the connotative system.

In many computational dual process models, (e.g., Freeman & Ambady, 2011; Glöckner & Betsch, 2008; Simon et al., 2015), both deliberative and intuitive systems are modeled as parallel constraint satisfaction (PCS). In such models, meaning arises dynamically as a result of the spreading or inhibition of activation between nodes representing single concepts connected in a small network. For example, the uncertain situation with a *nurse* or a *doctor* in a hospital setting could be modeled as mutually suppressing nodes representing a nurse or a doctor, respectively, with other information (such as the attribute *female*, the fact of wearing a white coat, or a powerful behavior) connected as activating (inhibiting)

https://doi.org/10.1017/S1930297500008652 Published online by Cambridge University Press

nodes to the compatible (incompatible) concept. As explained in detail by the PCS theory of stereotyping of Kunda & Thagard (1996), a simple iterative updating algorithm where activation is exchanged across all linked nodes until a stable configuration in the network is achieved, can model the holistic impression-formation process similar to the *BayesACT* example discussed in the previous section. Schröder & Thagard (2013) proposed a model of automatic social behavior that implements such a simple PCS network where the links between relevant concepts are empirically calibrated with the EPA dictionary databases also employed for the *BayesACT* model described here. However, these models are not dual-process models in a narrower sense as they provide either a denotative meaning structure set up rather arbitrarily (Kunda & Thagard, 1996) or a connotative meaning structure based on EPA studies (Schröder & Thagard, 2013), but not both.

The connectionist networks of Schröder & Thagard (2013) seek to build a formal model of the mutual influence of affect and cognition as a weight vector in a neural network. The Boltzmann machine is one possible mechanism to implement the weights in these connectionist networks, and minimizing this energy functional is the same as finding the posterior probability of the corresponding distribution over possible states (Ackley et al., 1985). This is the same process used for the somatic transform, implemented in *BayesACT* using a sample-based method (Hoey et al., 2016). The distributions in a Bayesian network could be implemented as one of many neural networks (see Aisa et al. (2008) for code and datasets); however, it is not obvious how the causal structure of the Bayesian network could be implemented.

Thagard (2006) introduced the HOTCO model (for "hot coherence"), a PCS network extended by a highly connected valence node and a valence parameter for every node representing a concept. This model can be regarded as a basic dual-process model with a rudimentary implementation of denotative (the conceptual structure of the PCS network) and connotative (the valence parameter of every node) meaning. However, unlike *BayesACT*, HOTCO is not rooted in a culturally shared meaning structure, so we would consider it an undersocialized model in Granovetter's sense (see above). Simon et al. (2015) give an in-depth discussion of HOTCO in the context of legal and moral decision-making, and apply the model to various scenarios. For example, passing judgment on some moral issue based on a number of facts that are manipulated to make experimental conditions. Their parallel constraint model includes "hot" cognitions such as "anger" and "liking", and shows how equilibria of the network correlate with findings of human behavior in experimental conditions. We discuss this work in greater detail in Section 5.3, where for comparison we map one of their scenarios to *BayesACT*.

Option and strategy choices are modeled in a dual parallel constraint network in Glöckner & Betsch (2008). The model has a "primary" network that rapidly settles to a maximally coherent description of the context and the revelation of an option to take, and a "secondary" network that is called upon if the primary network cannot find consistency. The secondary network then uses three strategies of information search, production and change, which are

https://doi.org/10.1017/S1930297500008652 Published online by Cambridge University Press

selected based on an evaluation of the extent to which it will help the primary network. However, exactly how this evaluation will be done is left open, reducing the precision of the proposed PCS model, showing it has some of the same issues as a multiple strategy theory. In *BayesACT*, the PCS of Glöckner & Betsch (2008) would include only denotative state, and would translate into a hierarchical Bayesian model in which the strategies map to the modifications of different parameters. Information production and change would be therefore shifts (potentially due to learning or experience) in the denotative transition dynamics in *BayesACT*. Information search would be modeled as another strategy of action, in which the goal is to reveal informative observables, much as in the active learning paradigm (Cohn et al., 1996) in the machine learning literature, and that could be optimally modeled in *BayesACT* as Bayesian reinforcement learning (BRL) (Duff, 2002) or connectionist models (Schmidhuber, 2013). POMDP policies may implement information seeking, encoding actions that refine beliefs (that provide estimates of risk), which are then used to choose more optimally. BRL takes this one step further and allows the POMDP to not only seek information about the state of the environment, but also about its own model.

A newer class of computational dual-process models related to PCS is inspired by the biological architecture of the human brain, simulating how intuitive versus deliberative judgments are computed by biological processes. For example, Ehret et al. (2015) proposed a neural-network implementation of multiple, iterative evaluative processes in hierarchically organized anatomical structures of the brain, from quick perceptual processing to complex semantic evaluations. Schröder et al. (2014) used the "semantic pointer" architecture of biological cognition (Eliasmith, 2013) to model the interplay of intentional and emotional action selection with a large neural-network model whose individual nodes do not represent semantic concepts but biologically realistic spiking neurons. Kajić et al. (2019) uses the same neurocomputational framework for a multi-level theory of emotions that integrates physiological processes, cognitive appraisals, and socio-cultural constructions. To account for social embeddedness of emotions, their models implement some of the datasets and the emotion model of affect control theory (see above), which are also at the core of the *BayesACT* model described here. There is thus a general conceptual compatibility of *BayesACT* with many of these neurocomputational dual-process models, but the focus is different. While these models attend to the neurobiological mechanisms underlying human judgment and decision-making, *BayesACT* targets the social embeddedness of these phenomena, while abstracting from neurological implementation details.

5 Exploratory Examples

Our aim in this section is to provide basic evidence for the generalizability of *BayesACT* as a model of human dual-process reasoning. The literature on human behavioral effects is vast, and we are only attempting to demonstrate how a single model could account for some effects across a range of different experimental setups. Our examples progress from the

https://doi.org/10.1017/S1930297500008652 Published online by Cambridge University Press

simple combinations of affective meanings with denotative representations and behaviors that explain fairness responses, through showing how the model can account for basic elements of cognitive dissonance, to an exploration of its relationship to coherence models based on parallel constraint satisfaction.

First, van den Bos (2001) showed how thoughts of uncertainty about the self can lead to more pro-social behavior. We show in Section 5.1 how this can be modeled as a tradeoff between denotative and connotative in affect control theory (ACT) and in *BayesACT*. Second, in a classic experiment, Festinger gave participants (teenage girls) one of two prizes of equal value to them (audio records of unknown pop stars) (Festinger, 1962). The participants subsequently raised their evaluations of the prize they obtained. In general, when a person is given one of two items that she values about the same, she will value the item she is given more highly in order to reduce the cognitive dissonance created by the fact that she did not get the other item. In Section 5.2, we show that, according to the somatic transform and *BayesACT*, such re-interpretation of value is simply the process of attempting to unify connotative representations of the self (e.g. "I am a good person") with denotative representations of uncertain events (e.g., "I think my prize is worse than hers"). Finally, in Section 5.3, we discuss how *BayesACT* can explain how moral decisions about guilt and innocence are modified by emotional factors such as sympathy, and we sketch how the parallel constraint model of this decision process by Simon et al. (2015) can be re-formulated as a *BayesACT* model. Parts of the fairness and PCS analyses are done initially using ACT alone, and the effects of *BayesACT* are explained after that. Throughout, we are presenting the minimum working example in order to demonstrate the use of the somatic transform in the modeling of cognitive biases.

5.1 Uncertainty and fairness

Van den Bos (2001) carried out an experiment in which the fairness or unfairness of a situation was evaluated affectively (positive vs. negative) in two conditions: one with induced (primed) feelings of uncertainty enhanced by asking about the emotions felt during uncertain episodes. Two effects are shown. First, the more (perceived) fair condition (where participants got to voice their opinion about a distribution of payoffs, but their opinions were ignored) elicited more positive emotions than the non-fair condition (no chance to voice). Second, the effect was *enhanced* by the increased uncertain feelings. In *BayesACT* this can be accounted for by noting that as the emotions associated with uncertainty about the self are evoked, so is the uncertainty in both denotative and connotative identity. However, as uncertainty about denotative identity is increased, the participant (a student) will be more reliant on the connotative system.

Consider a purely denotative solution. In this case, the emotions elicited will not play a role, and a student will think that voicing an opinion may change the payoffs in their favour, and so will prefer that option. However, they don't know if their voice will be taken into account, so the other option of letting the experimenter decide is only seen as marginally

https://doi.org/10.1017/S1930297500008652 Published online by Cambridge University Press

worse. On the other hand, a purely connotative system will focus on the emotional identity only, and so when feelings of uncertainty are evoked (uncertainty is made salient), a more negative emotional identity will react more negatively to the no-voice (default) condition, and the ability to voice opinion will make a more significant difference. This is because the voicing of opinion is an action with emotional meaning that restores feelings of certainty in identity.

We show the simplest possible example, using only non-probabilistic ACT equations in order to motivate the problem.[15] We show how *BayesACT* would modify things at the end of the section. In van den Bos (2001), there are 2×2 conditions, with half the participants having a "voice" and half not, and half of them having uncertainty made salient, half not. We therefore have two classes of participants: salient and non-salient. The non-salient identities are simply *student* (EPA:$\{1.5, 0.31, 0.75\}$), with a characteristic emotion of (EPA:$\{1.5, 0.66, 0.26\}$) with closest labels of *warm* and *easygoing*. The salient ones we interpret as *anxious student* (EPA:$\{-0.2, 0.04, 1.2\}$) using the modifier equations of ACT (see Section 2.2), with an emotion of (EPA:$\{-0.77, -0.30, 0.91\}$) with closest label of *envious*. These characteristic emotions are those the participants will report in the no-voice condition, so we see that the salient group will have more negative emotions.

In the voice condition, an action is taken by the participant, so we model the student taking the action *compromise with*, as this is close to the optimal for a student, and is what one would expect the student to do in the fairness test (divide equally). A *student* who *compromises with* another *student* feels emotions of (EPA:$\{1.9, 1.2, 0.029\}$) (closest label *patient*). On the other hand, if an *anxious student* is the actor, emotions are more positive (EPA:$\{2.2, 1.3, -0.15\}$) (closest label remains *patient*, however).

Figure 5(a) shows these data in a simple plot, where the "E" axis is reversed as in the experiments the participants were asked how "sad" and "disappointed" they were, thus a negative measure.We therefore plot the (inverted) scaled version (to the range $1-7$) of the distance between the "E" value of the emotion felt with the mean "E" value of the emotions *sad* (EPA:$\{-1.9, -1.7, -2.1\}$) and *disappointed* (EPA:$\{-1.7, 1.2, -1.3\}$). Thus, in the figure, larger y axis values correspond to more negative emotions. Note that these curves correspond in form to that observed in van den Bos (2001), shown in Figure 5(b). In essence, the "fairness" of a behavior is directly a function of how uncertain an agent is.

The ACT simulations shown in Figure 5 assume a participant is *either* uncertain (salient case) or not (non-salient case). However, in practice, there will be degrees of uncertainty, possibly based on individual personality (e.g. susceptibility to the probe generating the uncertainty). The simple affect control theory (ACT) model cannot handle this, but *BayesACT* explicitly trades off between the two cases continuously based on the amount of uncertainty induced. In the case of no uncertainty, the denotative solution takes over. As uncertainty grows, the connotative solution starts to dominate.

[15]Here we are using the EPA dataset compiled at Indiana University in 2005. Code to generate these results can be found at http://bayesact.ca.

https://doi.org/10.1017/S1930297500008652 Published online by Cambridge University Press

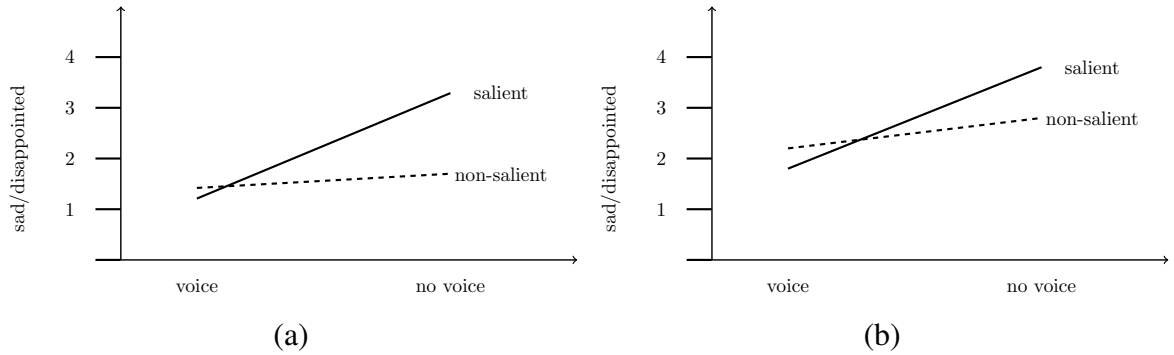


FIGURE 5: (a) ACT simulations of conditions, showing the scaled average of the distance from the emotion felt in the condition with the evaluation of the emotion of *sad* and *disappointed*. Scaling to the range 1-7 is done to match scales with that of van den Bos (2001). (b) Results of van den Bos (2001) showing the mean ratings of sadness and disappointment for each of the four cases. In both cases, results are shown as lines for exposition (data is 4 points: the line ends). Larger y axis values correspond to more negative emotions.

Thus, *BayesACT* can model how agents will be biased towards more connotative solutions by invoked feelings of anxiety leading to feelings of uncertainty. The correspondence between connotative predictions and experimental results of van den Bos (2001) show that people are leaning more heavily on the connotative meanings of the experiment in the uncertainty salient case. As uncertainty in the denotative identity is increased, the posterior over connotative identity becomes more focussed around the connotative meanings (of anxious student). If this were not the case, then the posterior would be more biased towards the denotative reality (of no control over outcomes, so a flat decision function), and the effect would not be as large. One can also see the same effects here as in FeldmanHall & Shenhav (2019), where uncertainty evoked negative affect, and restorative options for that affective state were preferred to restore affective meanings to something closer to their fundamental values. Lastly, this provides an indication of an experiment to falsify the model, as one could adjust the *degree* of uncertainty induced and observe the gradual transition across emotional states (essentially interpolating between the end-points in Figure 5).

5.2 Cognitive Dissonance

Cognitive dissonance is a broad term for a large literature of studies on how humans interpret and react to inconsistencies, and includes research on how inconsistency is identified, how affective dissonance is elicited, and how humans resolve inconsistencies (Gawronski & Brannon, 2019). There is growing consensus that the processes of dissonance and dissonance resolution has multiple facets, from the management of purely propositional (logical) inconsistencies (Gawronski & Brannon, 2019), some potentially action-oriented (Harmon-Jones & Harmon-Jones, 2018), to the relationship between the self and the situation (Stone & Cooper, 2001). Expected value has recently surfaced as a plausible interpretation that

https://doi.org/10.1017/S1930297500008652 Published online by Cambridge University Press

denies consistency its primacy for humans (Kruglanski et al., 2018). However, there is evidence that "values" are, in fact, exactly on those states that are expected, in the sense that human motivation is to work towards expected states (Friston, 2010), and deviation from expectations over the future create negative affect (Hesp et al., 2021). Thus, expected value and expectations over (denotative) futures may be one and the same, again returning consistency (over future expectations) to the center stage as a prime motivator for human action.

The self-standards model of Stone & Cooper (2001) makes a very similar point to ours about the self being used by humans as a reference mark from which to gauge inconsistent situations. Here we show that *BayesACT* is able to produce some of the same effects as accounted for by the self-standards model. While we work here on the free/forced choice paradigm, induced compliance and effort justification paradigms could be handled in a very similar way (Harmon-Jones & Harmon-Jones, 2018). Induced compliance with incentives (e.g. being forced to write a statement that contradicts moral beliefs, but being rewarded for it) can be modeled as a two-stage game where the positive affective meanings of the reward reduce negative affective feelings from the induced compliance. The belief disconfirmation paradigm, in which evidence is presented that contradicts a belief (Harmon-Jones & Harmon-Jones, 2018), is modeled as any transient impression in *BayesACT*.

BayesACT, as a multi-attribute Bayesian decision network, can in theory also represent any set of propositions, and thus can be used to detect inconsistencies by looking at joint probabilities of related variables, and can compute expected value of outcomes with both consistent and inconsistent beliefs (Kruglanski et al., 2018). Any such inconsistencies, however, will increase the entropy (dispersion) for the denotative system in *BayesACT*, and lead to an increasing reliance on connotative interpretations. As a result, denotative action choice distributions are less informative, making behavior initiation more difficult. The induced inconsistencies in the connotative system will be felt as dissonance, often (but not always) as negative emotion. The Bayesian model naturally trades off between the two. As we show in this section, for a simple (and highly stylized) experimental setup, *BayesACT* shows some of the basic properties of a self-standards type model of cognitive dissonance, combined with those of a propositional inconsistency and expected values model, and may offer a unifying view that combines these different explanations.

Consider a simple demonstrative example in which X corresponds to whether an item is desirable ($X = good$) or not ($X = bad$). The corresponding Y (given by $y = M(X = x)$) is the "E" rating for the item. As demonstrated by Shank & Lulham (2017), people are consistent in their ratings of the EPA values of commercial products. For example, iPhones were rated as (EPA:$\{1.3, 1.0, 1.5\}$), whereas Blackberry phones were rated (EPA:$\{-0.67, -0.71, -0.28\}$). The study also found that commercial products change people's identities, and are seen as consistent with some identities and not with others. For example, a CEO with an iPhone is perceived as more powerful as one without. We place a prior on Y that is the same as the identity of the participant.

https://doi.org/10.1017/S1930297500008652 Published online by Cambridge University Press

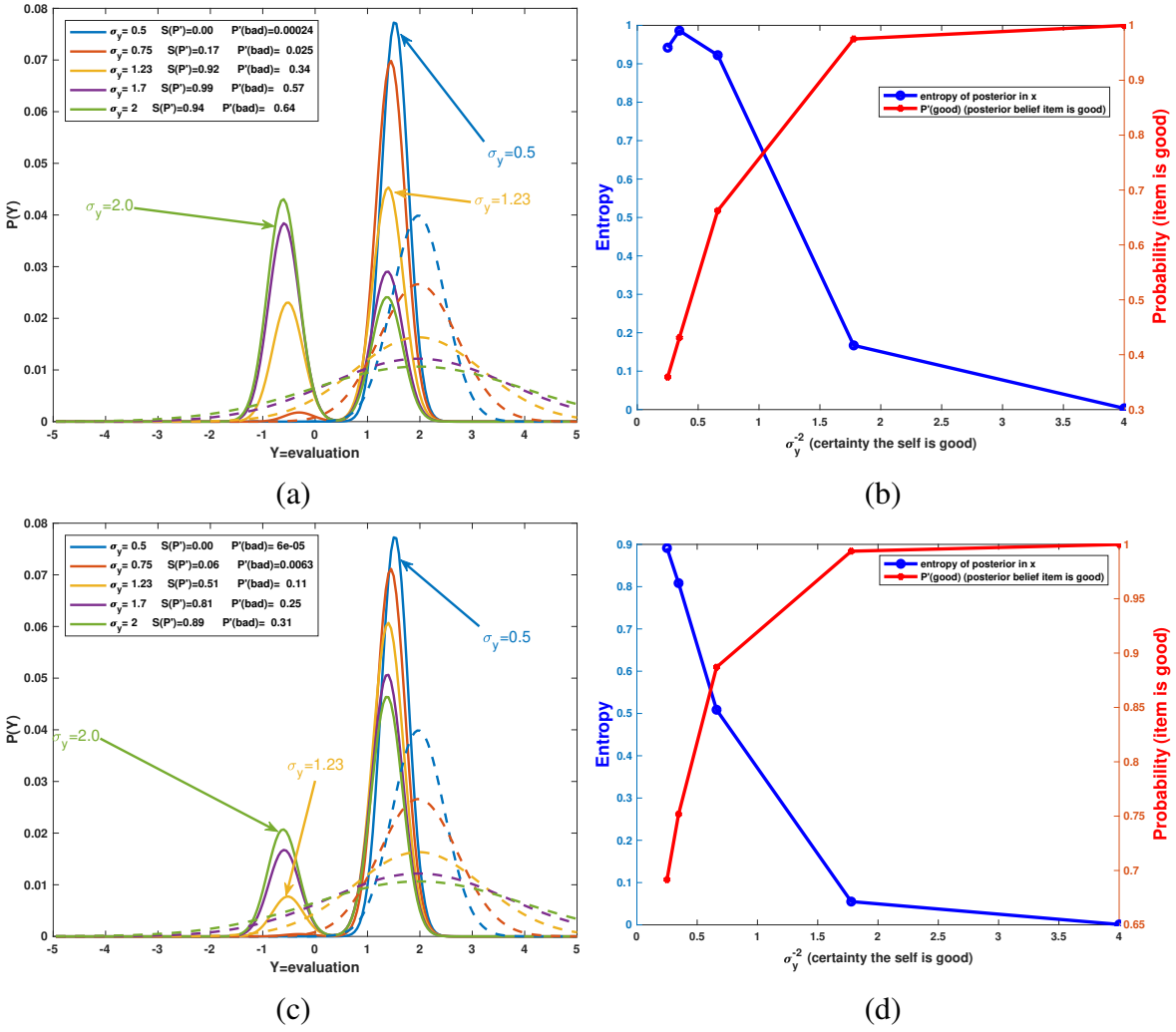


FIGURE 6: Simulation of a cognitive dissonance. (a) The posteriors over X and Y shift towards the prior over Y in a forced-choice paradigm, causing a re-interpretation of a *bad* item as something *good*. The prior in Y has a stronger effect if it is less dispersed (smaller σ_y, dashed lines). $S(P')$ is the entropy of $P'(X)$ and $P'(bad)$ is the posterior probability of $X = bad$. The prior in X is $P(X = bad) = 0.8$. (b) corresponding plot of the overall trends in entropy and posterior probability as a function of σ_y^{-2} (certainty the self is "good"). (c)-(d) same as (a)-(b) but for prior $P(X = bad) = 0.5$ (free choice paradigm).

In Figure 6, the y axis corresponds to the evaluative dimension "E", and the prior $P(y)$ has $\mu_y = 2.0$ and $\sigma_y = 1.23$ (corresponding to the mean and standard deviation of the E rating for *child* in the University of Georgia 2015 dataset).[16]

We used $\gamma = 0.3$ and imagine an experiment where the participant is given a Blackberry.

[16]We selected *child* for this demonstrative example because its evaluation and potency are more consistent across decades than *teenager* when comparing the 2005 data with the 2015 data. As we are trying to replicate experiments from the 1970s with young adults, the term "child" may be more appropriate and general. A more in-depth examination of this process would include a direct measurement of sentiment about the self of the participants.

https://doi.org/10.1017/S1930297500008652 Published online by Cambridge University Press

The denotative prior is $P(X = bad) = 0.8$, implying the participant believes they have a bad item. After combining the connotative prior (which is essentially saying that any item obtained by the participant must be *good*, since they are *good* and expect to have *good* things), the resulting posterior has a reduced value for $P'(X = bad)$ (dropping to 0.34), so is significantly more likely to be on the *good* side. That is, a participant who originally thought the prize was not as good ($P(X = bad) = 0.8$), has changed her or his mind and now thinks the prize is much better ($P'(X = bad) = 0.34$).

Figure 6(a)-(b) also shows the posteriors for a range of smaller (0.5) to larger (2.0) values of σ_y. With a more dispersed prior (larger σ_y), the shift is not as evident ($P'(X = bad) = 0.64$), and with a less dispersed prior (smaller σ_y), the shift is even more evident ($P'(X = bad) = 0.00024$). *BayesACT* predicts that agents will deal with less valid environments by leaning more heavily on their connotative system. Thus, the resulting $P'(x)$ is low precisely because the connotative system has "taken over" and it has become more imperative to justify receiving the lesser gift.

Any actual experiment would need to take a range of "types" of σ_y into account by measuring the distribution over σ_y in the population and then integrating them out as:

$$P(X = bad) = \int_{j \in \text{types of } \sigma_y} P(\sigma_y = j) P(X = bad | \sigma_y = j). \tag{7}$$

where $P(\sigma_y = j)$ is the empirical prior for the population under study.

Figure 6(c)-(d) shows the same simulation but this time with $P(X = bad) = 0.5$. With this prior, either item is equivalent from an expected utility standpoint, and so an optimal denotative strategy is to choose randomly, thus mirroring the free choice paradigm (Brehm, 1956) in which two equally desirable items are to be chosen between. The results are very similar here, with the obtained (chosen) item being evaluated much more positively (even more so than in the forced choice case considered above).

In the self-standards model (Stone & Cooper, 2001), a focus on either personal attributes of the self, or societal norms can both create dissonance, as can non-self related inconsistencies. In the *BayesACT* view, societal norms are implicitly embedded in the self-sentiment, and so the first two aspects are really considered to be the same when viewed through the connotative lens of affect control theory. Logical inconsistencies (Gawronski & Brannon, 2019) can also be accounted for in *BayesACT*, although we have not explored this in detail here. However, changes in expected value certainly can, but we argue that affective reactions to changes in expected (denotative) value are tempered by affective reactions to changes in (connotative) meanings about the self. Thus, a student who expects to do badly on a test is confronted with inconsistency if he does well, but this is argued by Kruglanski et al. (2018) not to create negative affect as predicted by classical dissonance theory. However, as pointed out by Harmon-Jones & Harmon-Jones (2018), other elements (such as self-concept changes) will also play a role, possibly overriding the negative affective reactions to inconsistency. In *BayesACT*, these effects are handled as changes in denotative state (from doing badly on the test, to doing well), increase in denotative entropy as a result, and

https://doi.org/10.1017/S1930297500008652 Published online by Cambridge University Press

thereby increased salience of connotative meanings (shifting from "bad student" to "good student"). Similarly, denotative inconsistencies that have weak connotative meanings (e.g. drawing colored balls from a box (Kruglanski et al., 2018)) would not change the posterior distribution over the connotative state, and so would not generate any non-characteristic affect, as observed in empirical results (Kruglanski et al., 2018).

As a unifying view, *BayesACT* shows that the negative affect is less related to inconsistency or value *per se*, but rather to the effects that this has on the connotative state, modulated by uncertainty. For example, *BayesACT* would predict someone with a strong negative self-concept (who thinks of themselves as a "bad student"), will discount the positive test score instead, possibly arriving at a denotative interpretation that they did badly in any case (regardless of the evidence to the contrary). This means cognitive dissonance experiments that explicitly queried participants for their identity would yield falsifiable *BayesACT* predictions such as this. Many of the disagreements in cognitive dissonance theories could be resolved by considering that there are *two* types of inconsistency which are tightly related: dissonant cognitions (or logical inconsistencies), and dissonant feelings (invalidation of self-concepts). This mirrors Festinger's original proposal of two determinants of dissonance: cognition and importance to the individual (Festinger, 1962). Dissonant cognitions (or incoherence between denotative priors and evidence in *BayesACT*) do not elicit affect directly in *BayesACT*, but they increase entropy leading to increased reliance on connotative meanings, which are the source of affective reactions.

One key element of the *BayesACT* simulations described above is that the certainty in the self-identity is highly relevant for the amount of dissonance caused. In the cognitive dissonance experiment described, the variance of the self-identity is coupled with the act of "owning" something. In the slightly broader category of cognitive biases normally labeled as self-affirmation (Sherman & Cohen, 2006), similar styles of experiments have shown how people will be more objective in the analysis of facts when self-affirmed. To put this in the same language as the cognitive dissonance experiments, they are more willing to re-interpret facts that disagree with their values in a way that changes their view on those facts. The increased dispersion in the denotative posterior allows an agent to entertain scenarios that disagree with their values.

Confirmation biases are another form of the same effect. This is the observation that people's prime concern is what others think of them, and that this drives them to seek evidence that agrees with their beliefs. In *BayesACT*, this concern is modeled in the connotative state, and reasoning is used to construct a viable denotative trajectory that leads to the connotative prescription. That is the difference between *should* and *would* (Patterson, 2014; Martin & Lembo, 2020), or between *must I* (reason) and *can I* (intuition) (Haidt, 2012).

https://doi.org/10.1017/S1930297500008652 Published online by Cambridge University Press

5.3 Probabilistic Constraint Networks in *BayesACT*

In this section, we briefly sketch how to map a connectionist parallel constraint satisfaction (PCS) model to a *BayesACT* model. Our goal is to outline how the two types of model are largely compatible. The primary difference between them is in the learning bias they apply, and so the selection of one over the other would come down to a much more detailed analysis of which bias is more successful in generalizing across situations, an empirical question yet to be answered.

We base this analysis on a recent experiment by Simon et al. (2015) where a participant is asked to judge a student named "Debbie" who is accused of cheating, based on a number of facts that are manipulated to make experimental conditions. The participants are assigned the role of *Judicial Officer* in charge of deciding on whether to suspend the student or not. A key element in such a PCS model is a pair of inhibitory nodes that model complementary concepts, such as whether Debbie is guilty or not. In *BayesACT*, each such pair is described by a single random variable with two states (e.g. a binary variable called "guilty" in the set $\mathbf{X}$ with values "true" and "false"). The probability distribution over these two states maps to the "activation" of the pair of nodes in the PCS. For example, suppose the activation of the "not guilty" and "guilty" nodes in the PCS are 0.72 and 0.45. This could map to a probability distribution $P(guilty) \propto [e^{0.72}, e^{0.45}] \propto [0.58, 0.43]$ over the "guilty" random variable in *BayesACT*. Each pair of guilt and innocence facts are modeled with binary observation variables in *BayesACT*, and each has a higher probability given its corresponding value of the state. That is, $P(guilt\ fact|guilty) > P(guilt\ fact|innocent)$ and $P(innocent\ fact|innocent) > P(innocent\ fact|guilty)$. The emotional nodes such as "sympathy" would be modeled as identity traits in *BayesACT*, such as *sympathetic judicial officer*, and in the somatic transform linking positive valence and "liking" with the identities assumed in the interaction (e.g. the participant might "like" Debbie because they have a friend named Debbie). Motivation leading to action in the PCS is modeled with a preference or utility function that ranks outcomes in *BayesACT*. In contrast to the model by Simon et al. (2015), the described mappings are not designed specifically by the researcher, but taken from empirical EPA databases that reflect culturally shared connotative meaning structures (Ambrasat et al., 2014; Heise, 2010).

Decision theoretically, the *BayesACT* POMDP will select the "correct" action (convict the guilty and acquit the innocent), reducing the problem to one of inference about the guilty/innocent denotative state variable. Nevertheless, *BayesACT* explicitly includes a variable representing action that is deeply connected to its affective meaning through the somatic transform. Further, the identities selected for the two persons being modeled will be based upon the text they read, and would thus provide a fast estimate of a probability distribution over the affective space indicating how the best behavior should "feel". This distribution then weights the denotative action distribution, increasing the likelihood of the agent taking the action that is affectively more aligned with this prediction.

To get a better feel for the identity dynamics involved, suppose we choose the identi-

https://doi.org/10.1017/S1930297500008652 Published online by Cambridge University Press

TABLE 1: Deflections for different conditions where the juror or friend (actor) can be sympathetic or not and the object (client) can be a student, delinquent or friend. The lowest deflection for each actor-object pair is shown in bold.

		actor identity			
		juror		friend	
behavior	client identity	non-sympathetic	sympathetic	non-sympathetic	sympathetic
convict	student	**1.7**	2.4	4.9	4.8
forgive	student	2.0	**1.6**	**0.8**	**0.6**
convict	delinquent	**1.1**	**1.9**	**4.7**	**4.5**
forgive	delinquent	4.8	5.1	6.7	6.5
convict	friend	4.2	4.8	6.9	6.8
forgive	friend	**3.0**	**2.5**	**1.5**	**1.4**

ties of *juror* (EPA:$\{0.66, 1.3, 0.060\}$)[17] and *student* (EPA:$\{1.5, 0.31, 0.75\}$). The identity of the *student* is fluid, however, because it depends (through the somatic transform) on the denotative state variable of guilt/innocence. That is, we will have a prior belief that students will be more likely to be innocent than guilty, rooted in culturally shared connotations of the concept *student*. We also consider two other identities for Debbie: *friend* (EPA:$\{2.8, 1.9, 1.4\}$) and *delinquent* (EPA:$\{-1.8, -0.78, 0.41\}$). Considering the deflections (emotional incoherence) of an ACT simulation of this situation, we use behaviors of *convict* (EPA:$\{0.05, 1.36, 0.34\}$) and *forgive* (EPA:$\{2.6, 2.0, 0.40\}$).[18] Further, we adjust the juror identity to *sympathetic juror* using modifier equations to model the condition in which the sympathy is induced, giving a rating of *sympathetic juror* of (EPA:$\{1.3, 1.1, -0.34\}$).

Considering the first two columns in Table 1, the deflection caused by forgiving is lower than convicting in the sympathetic juror case, while it is higher in the non-sympathetic juror case. On average therefore, in agreement with the results of Simon et al. (2015), there will be more acquittals in the sympathy case because of this effect. Should the juror consider the student to be a *delinquent*, on the other hand, then the more likely behaviour will be to convict, regardless of sympathy. Finally, forgiveness is more likely if Debbie is considered a *friend*. Although three pairs of identities does not give conclusive evidence, it is certainly evidence that the *BayesACT* model would show the same effects when taken in population averages. Participants may not all assign identities of student or delinquent to Debbie, and

[17]Data in this section taken from the Indiana University 2005 EPA survey.

[18]Both this choice and that of *juror* are because *acquit* and *Judicial Officer* are not in the Indiana 2005 dataset so synonyms were chosen. The choice of the exact concept is not important, as the *BayesACT* simulation would represent a distribution over these concepts, the shape of which could be controlled as independent variable in an experiment.

https://doi.org/10.1017/S1930297500008652 Published online by Cambridge University Press

may not all assign identities of juror to themselves. However, we know that, on average, people will consider cheaters to be more negative and powerless identities, and so the same effect will show up and would be averaged over in *BayesACT*.

As stated earlier, this example gives a simple mechanism for falsifying *BayesACT*, as *BayesACT* predicts that if we change the participant's assumed role (identity), the final decisions would change. The model of Simon et al. (2015) would make the same prediction unless all the weights were modified to reflect this or a new node was added to the coherence network. Table 1 (last two columns) further shows what the deflections look like if the situation was re-framed and the participants were told to imagine they were a *friend* of Debbie's. In this case, the deflections are generally higher for convicting, because a *friend* is inherently relational and somewhat expects the object-person to be a *friend* as well, so dealing (as *friend*) with a *student* or *delinquent* is more deflecting (less likely). Further, the best action to take is always to forgive if the object is a *student* or a *friend*, regardless of sympathy.

While a new node for the identity could be added to the Simon *et al.* model, it would need to be precisely tailored for the domain, and connections added with weights that would determine that friends don't convict students, for example. *BayesACT* is more general because one needs only to change the identity, and the same model can be used. Further, uncertainty in *BayesACT* allows one to modify the experimental condition making it more or less ambiguous, and this can be explicitly taken into account. *BayesACT* predicts that making the facts less ambiguous would reduce the effect of the connotative network and thus of the identity labels.

6 Conclusion

In this paper, we proposed *BayesACT* as a computational dual-process model of human interactions, and showed how it explicitly represents a tradeoff between the uncertainty in a denotative space (of e.g. symbolic constructs about the physics of the world) and in a connotative space (of feelings about identities and behaviors). We argued that *BayesACT* captures some of the key elements of known human dual-process reasoning, while embedding decision-making in a socio-cultural environment.

We reviewed some of the connections between *BayesACT* and parallel constraint models, including those implemented as neural networks. There is evidence now that the two can be translated into one another (Rohekar et al., 2018), but this remains as future work. Other future avenues for work include the more in-depth investigation of planning methods, likely starting from some of the insights in Asghar & Hoey (2015), in which Monte-Carlo tree search plays an instrumental role. Finally, scaling the model to more complex domains both connotatively and denotatively is a priority.

We have attempted to argue that *BayesACT* is an abstraction that "fits" the evidence we have (from surveys) to a certain degree. However, this does not imply that there are

https://doi.org/10.1017/S1930297500008652 Published online by Cambridge University Press

not other factors at play, including other connotative ones. As in any complex system, the interactions between these other factors and the emotional coherence modeled in *BayesACT* may have highly non-linear or emergent effects that are difficult to predict. However, *ceteris paribus*, we believe the effects shown here may provide a level of generality that is useful for the modeling of human behavior.

We discussed how uncertainty plays a critical role in determining the relative contributions of denotative and connotative reasoning, with more uncertainty leading to action choices more in line with connotative (affective) meanings, while less uncertainty engenders more deliberative (denotative) action choices. We discussed the relationship of *BayesACT* to other dual process theories and to other social psychological and sociological theorizing. Finally, we demonstrated how this simple idea can be used to generate reasonable solutions to well-known human cognitive biases. Our objective is to continue applying the model to other cognitive biases to explore the limits of its generalizability. We are currently investigating occupational status before and after the COVID-19 pandemic, replicating Freeland & Hoey (2018), and we also plan to use it to study small group and job management processes online. Other potential applications include relationship therapy, job performance evaluation, and intelligent tutoring.

References

Åström, K. J. (1965). Optimal Control of Markov Decision Processes with Incomplete State Estimation. *J. Math. Anal. App.*, *10*, 174–205.

Ackley, D. H., Hinton, G. E., & Sejnowski, T. J. (1985). A learning algorithm for Boltzmann machines. *Cognitive science*, *9*(1), 147–169.

Aisa, B., Mingus, B., & O'Reilly, R. (2008). The Emergent neural modeling system. *Neural networks*, *21*(8), 1146–1152.

Akerlof, G. A. & Kranton, R. E. (2000). Economics and Identity. *The Quarterly Journal of Economics*, *115*(3), 715–753.

Ambrasat, J., von Scheve, C., Conrad, M., Schauenburg, G., & Schröder, T. (2014). Consensus and stratification in the affective meaning of human sociality. *Proceedings of the National Academy of Sciences*, *111*(22), 8001–8006.

Ambrasat, J., von Scheve, C., Schauenburg, G., Conrad, M., & Schröder, T. (2016). Unpacking the Habitus: Meaning Making Across Lifestyles. *Sociological Forum*, *31*(4), 994–1017.

Asghar, N. & Hoey, J. (2015). Monte-Carlo Planning for Socially Aligned Agents using Bayesian Affect Control Theory. In *Proc. Uncertainty in Artificial Intelligence (UAI)* (pp. 72–81).

Bail, C. A., et al. (2018). Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, *115*(37), 9216–9221.

https://doi.org/10.1017/S1930297500008652 Published online by Cambridge University Press

Barrett, L. F. & Satpute, A. (2013). Large-scale Brain Networks in Affective and Social Neuroscience: Towards an Integrative Functional Architecture of the Brain. *Current Opinion in Neurobiology*, *23*, 361–372.

Berger, P. L. & Luckmann, T. (1967). *The social construction of reality*. Doubleday: Anchor Books.

Blumer, H. (1969). *Symbolic interactionism: perspective and method*. University of California Press.

Bourdieu, P. (1990). *The logic of practice*. Stanford University Press.

Boutilier, C., Dean, T., & Hanks, S. (1999). Decision Theoretic Planning: Structural Assumptions and Computational Leverage. *Journal of Artifical Intelligence Research*, *11*, 1–94.

Brehm, J. (1956). Postdecision changes in the desirability of alternatives. *Journal of Abnormal and Social Psychology*, *52*, 384–389.

Bruch, E. & Feinberg, F. (2017). Decision-Making Processes in Social Contexts. *Annual Review of Sociology*, *43*, 207–227.

Capraro, V. & Rand, D. G. (2018). Do the Right Thing: Preferences for Moral Behavior, Rather Than Equity or Efficiency per se, Drive Human Prosociality. *Judgment and Decision Making*, *13*(1), 99–111.

Chaiken, S. & Trope, Y. (1999). *Dual-process theories in social psychology*. New York: Guildford.

Cohn, D. A., Ghahramani, Z., & Jordan, M. I. (1996). Active Learning with Statistical Models. *Journal of Artificial Intelligence Research*, *4*, 129–145.

Conant, R. C. & Ashby, W. R. (1970). Every good regulator of a system must be a model of that system. *International Journal of Systems Science*, *1*(2), 89–97.

Correll, S. J., Weisshaar, K. R., Wynn, A. T., & Wehner, J. D. (2020). Inside the Black Box of Organizational Life: The Gendered Language of Performance Assessment. *American Sociological Review*, *85*(6), 1022–1050.

Doshi-Velez, F. (2009). The infinite partially observable Markov decision process. In *Advances in neural information processing systems* (pp. 477–485).

Duff, M. O. (2002). *Optimal Learning: Computational procedures for Bayes-adaptive Markov decision processes*. PhD thesis, University of Massassachusetts Amherst.

Duncan, S. & Barrett, L. F. (2007). Affect is a Form of Cognition: A Neurobiological Analysis. *Cognition and Emotion*, *21*, 1184–1211.

Ehret, P. J., Monroe, B. M., & Read, S. J. (2015). Modeling the dynamics of evaluation: A multilevel neural network implementation of the iterative reprocessing model. *Personality and Social Psychology Review*, *19*(2), 148–176.

Eliasmith, C. (2013). *How to build a brain: a neural architecture for biological cognition*. New York, NY: Oxford University Press.

Evans, J. S. (2008). Dual-Processing Accounts of Reasoning, Judgment and Social Cognition. *Annu. Rev. Psychol.*, *59*, 255–278.

Fehr, E. & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics*, *114*(3), 817–868.

FeldmanHall, O. & Shenhav, A. (2019). Resolving uncertainty in a social world. *Nature Human Behaviour*, *3*, 426–435.

Festinger, L. (1962). Cognitive dissonance. *Scientific American*, *207*(4), 93–107.

Forgas, J. P. (2008). Affect and Cognition. *Perspectives on Psychological Science*, *3*, 94–101.

Freeland, R. & Hoey, J. (2018). The Structure of Deference: Modeling Occupational Status Using Affect Control Theory. *American Sociological Review*, *83*(2), 243–277.

Freeman, J. B. & Ambady, N. (2011). A dynamic interactive theory of person construal. *Psychological Review*, *118*(2), 247–279.

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

Gawronski, B. & Brannon, S. M. (2019). What is Cognitive Consistency and Why Does It Matter? In E. Harmon-Jones (Ed.), *Cognitive dissonance: progress on a pivotal theory in social psychology*. Washington DC: American Psychological Association, second edition.

Glöckner, A. & Betsch, T. (2008). Modeling Option and Strategy Choice with Connectionist Networks: Towards and Integrative Model of Automatic and Deliberate Decision Making. *Judgment and Decision Making*, *3*(3), 215–228.

Glöckner, A. & Betsch, T. (2011). The empirical content of theories in judgement and decision making: Shortcomings and remedies. *Judgment and Decision Making*, *6*(8), 711–721.

Glöckner, A. & Witteman, C. (2009). Beyond dual-process models: a categorization of processes underlying intuitive judgement and decision making. *Thinking and Reasoning*, *16*(1), 1–25.

Granovetter, M. (1985). Economic Action and Social Structure: The Problem of Embeddedness. *American Journal of Sociology*, *91*(3), 481–510.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, *108*, 814–834.

Haidt, J. (2012). *The righteous mind: why good people are divided by politics and religion*. New York: Vintage.

Harmon-Jones, C. & Harmon-Jones, E. (2018). Toward an Increased Understanding of Dissonance Processes: A Response to the Target Article by Kruglanski et. al. *Psychological Inquiry*, *29*(2), 74–81.

Heise, D. R. (2007). *Expressive order: confirming sentiments in social actions*. Springer.

Heise, D. R. (2010). *Surveying cultures: discovering shared conceptions and sentiments*. Wiley.

Hesp, C., Smith, R., Allen, M., Friston, K., & Ramstead, M. (2021). Deeply Felt Affect: The Emergence of Valence in Deep Active Inference. *Neural computation*, *33*(2), 398–446.

Hoey, J. & Schröder, T. (2015). Bayesian Affect Control Theory of Self. In *Proceedings of*

the AAAI Conference on Artificial Intelligence (pp. 529–536).

Hoey, J., Schröder, T., & Alhothali, A. (2016). Affect Control Processes: Intelligent Affective Interaction using a Partially Observable Markov Decision Process. *Artificial Intelligence*, *230*, 134–172.

Jaini, P. & Poupart, P. (2016). Online and distributed learning of Gaussian mixture models by Bayesian moment matching. *arXiv preprint arXiv:1609.05881*.

Kaelbling, L. P., Littman, M. L., & Cassandra, A. R. (1998). Planning and Acting in Partially Observable Stochastic Domains. *Artificial Intelligence*, *101*, 99–134.

Kahan, D. M. (2008). Cultural cognition as a conception of the cultural theory of risk. In S. Roeser (Ed.), *Handbook of risk theory* (pp. 08–20). Springer.

Kahneman, D. & Klein, G. (2009). Conditions for intuitive expertise: A failure to disagree. *American Psychologist*, *64*(6), 515–526.

Kajić, I., Schröder, T., Stewart, T. C., & Thagard, P. (2019). The semantic pointer theory of emotion: Integrating physiology, appraisal, and construction. *Cognitive Systems Research*, *58*, 35–53.

Koller, D. & Friedman, N. (2009). *Probabilistic graphical models: principles and techniques*. MIT Press.

Kruglanski, A. W., Jasko, K., Milyavsky, M., Chernikova, M., Webber, D., Pierro, A., & di Santo, D. (2018). Cognitive Consistency Theory in Social Psychology: A Paradigm Reconsidered. *Psychological Inquiry*, *29*(2), 45–59.

Kunda, Z. & Thagard, P. (1996). Forming impressions from stereotypes, traits, and behaviors: A parallel-constraint-satisfaction theory. *Psychological Review*, *103*(2), 284–308.

Lange, C. G. & James, W. (1922). *The emotions*. New York: Hafner.

Lazarus, R. (1984). On the Primacy of Cognition. *American Psychologist*, *39*, 124–129.

MacKinnon, N. J. (1994). *Symbolic interactionism as affect control*. Albany: State University of New York Press.

MacKinnon, N. J. & Heise, D. R. (2010). *Self, identity and social institutions*. New York, NY: Palgrave and Macmillan.

MacKinnon, N. J. & Hoey, J. (2021). Operationalizing the Relation between Affect and Cognition with The Somatic Transform. *Article under review*.

Martin, J. L. & Lembo, A. (2020). On the Other Side of Values. *American Journal of Sociology*, *156*(1), 52–98.

Mather, M. & Fanselow, M. S. (2018). Editorial overview: Interactions between Emotion and Cognition. *Current Opinion in Behavioral Sciences*, *19*, iv – vi.

Mead, G. H. (1934). *Mind, self and society*. University of Chicago Press.

Messick, D. M. & McClintock, C. G. (1968). Motivational Bases of Choice in Experimental Games. *Journal of Experimental Social Psychology*, *4*, 1–25.

Mook, D. G. (1987). *Motivation: the organization of action*. New York: Norton.

Ortony, A., Norman, D., & Revelle, W. (2005). Affect and proto-affect in effective functioning. In J. Fellous & M. Arbib (Eds.), *Who needs emotions: the brain meets the machine*

(pp. 173–202). Oxford University Press.

Osgood, C. E. (1969). On the Whys and Wherefores of EPA. *Journal of Personality and Social Psychology*, *12*, 194–199.

Osgood, C. E., May, W. H., & Miron, M. S. (1975). *Cross-cultural universals of affective meaning*. University of Illinois Press.

Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). *The measurement of meaning*. Urbana: University of Illinois Press.

Patterson, O. (2014). Making Sense of Culture. *Annual Review of Sociology*, *40*, 1–30.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. San Mateo, CA: Morgan Kaufmann.

Puterman, M. L. (1994). *Markov decision processes: discrete stochastic dynamic programming*. New York, NY.: Wiley.

Rabin, M. (1993). A theory of fairness, competition and cooperation. *The American Economic Review*, *83*(5), 1281–1302.

Rohekar, R. Y., Nisimov, S., Gurwicz, Y., Koren, G., & Novik, G. (2018). Constructing deep neural networks by Bayesian network structure learning. In *Advances in Neural Information Processing Systems* (pp. 3047–3058).

Russell, S. & Norvig, P. (2010). *Artificial intelligence: a modern approach*. Prentice Hall, third edition.

Schmidhuber, J. (2013). POWERPLAY: Training an Increasingly General Problem Solver by Continually Searching for the Simplest Still Unsolvable Problem. *Frontiers in Psychology*, *4*(313).

Schröder, T., Hoey, J., & Rogers, K. B. (2016). Modeling Dynamic Identities and Uncertainty in Social Interactions: Bayesian Affect Control Theory. *American Sociological Review*, *81*(4), 828–855.

Schröder, T., Stewart, T. C., & Thagard, P. (2014). Intention, emotion, and action: A neural theory based on semantic pointers. *Cognitive Science*, *38*(5), 851–880.

Schröder, T. & Thagard, P. (2013). The affective meanings of automatic social behaviors: Three mechanisms that explain priming. *Psychological Review*, *120*, 255–280.

Shank, D. B. & Lulham, R. (2017). Products as Affective Modifiers of Identities. *Sociological Perspectives*, *60*(1), 186–205.

Sherman, D. K. & Cohen, G. L. (2006). The Psychology of Self-Defense: Self-Affirmation Theory. *Advances in Experimental Social Psychology*, *38*, 183–242.

Shiller, R. J. (2017). Narrative economics. *American Economic Review*, *107*(4), 967–1004.

Silver, D. & Veness, J. (2010). Monte-Carlo Planning in Large POMDPs. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, & A. Culotta (Eds.), *Advances in Neural Information Processing Systems (NIPS) 23* (pp. 2164–2172). Curran Associates, Inc.

Simon, D., Stenstrom, D. M., & Read, S. J. (2015). The Coherence Effect: Blending Cold and Hot Cognitions. *Journal of Personality and Social Psychology*, *109*(3), 369–394.

Smith, R., Parr, T., & Friston, K. J. (2019). Simulating Emotions: An Active Inference

Model of Emotional State Inference and Emotion Concept Learning. *Frontiers in Psychology*, *10*, 2844.

Stanovich, K. E. & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences*, *23*, 645–726.

Stone, J. & Cooper, J. (2001). A Self-Standards Model of Cognitive Dissonance. *Journal of Experimental Social Psychology*, *37*, 228–243.

Thagard, P. (2006). *Hot thought: mechanisms and applications of emotional cognition*. MIT Press.

Vaisey, S. & Valentino, L. (2018). Culture and choice: Toward integrating cultural sociology with the judgment and decision-making sciences. *Poetics*, *68*, 131 – 143.

van den Bos, K. (2001). Uncertainty management: The influence of uncertainty salience on reactions to perceived procedural fairness. *Journal of Personality and Social Psychology*, *80*(6), 931–941.

Vives, M. L. & FeldmanHall, O. (2018). Tolerance to ambiguous uncertainty predicts prosocial behaviour. *Nature Communications*, *9*(2156).

von Neumann, J. & Morgenstern, O. (1953). *Theory of games and economic behavior*. Princeton, NJ: Princeton University Press, 3 edition.

Wood, W., Quinn, J., & Kashy, D. (2002). Habits in everyday life: Thought, emotion, and action. *Journal of Personality and Social Psychology*, *83*(6), 1281–1297.

Zajonc, R. (1980). Feeling and Thinking: Preferences Need No Inferences. *American Psychologist*, *35*, 151–175.

Zajonc, R. (1984). On the Primacy of Affect. *American Psychologist*, *39*, 117–123.

Appendix

Abstractions and simplifications from full *BayesACT*

In this section, we relate the full *BayesACT* model to the presentation of the "simplified" model we are using to demonstrate the use of the somatic transform. In that presentation, we essentially aggregate or abstract away all elements of the model that are not directly connected through the somatic transform in order to focus exclusively on that aspect. However, when using *BayesACT* in full, temporal dynamics and observations need to be accounted for.

The posterior $P(X', Y')$ that we are exploring in Section 3 is computed by starting from a prior over denotative and connotative state factored as $P(X, Y) = P(X)P(Y)$. This factorization can be performed because the somatic transform links the two as described by Equation 2. We write this distribution as a belief function $b(x, y) = b(x)b(y)$ in the following:

$$P(x', y'|G', G, b(x, y), \omega'_e, \omega'_x) \propto G'(x', y')P(x'|b(x), G, \omega'_x)P(y'|b(y), G, \omega'_e) \quad (8)$$

$$\propto G(x', y')P(\omega'_x|x', b(x), G)P(x'|b(x), G)$$

$$P(\omega'_e|y', b(y), G)P(y'|b(y), G) \quad (9)$$

$$= G(x', y') \sum_x P(\omega'_x|x')P(x'|b(x), G) \int_y P(\omega'_e|y')P(y'|b(y), G) \quad (10)$$

$$= G(x', y') \Big[\sum_x P(\omega'_x|x')P(x'|b(x)) \Big] \Big[\int_y P(\omega'_e|y')P(y'|b(y)) \Big] \quad (11)$$

In Section 3, we are manually specifying the two terms $[\ldots]$, and considering them as priors $P(x)$ and $P(y)$ so that we can directly examine the effects of the somatic transform, G. That is, in Equations 3 and 4, we are manually specifying $P(x)$ and $P(y)$, while in a sequential simulation, these terms would be given as:

$$P(x) : \sum_x P(x'b(x)|\omega'_x) \propto \sum_x P(\omega'_x|x')P(x'|x)b(x), \quad (12)$$

and

$$P(y) : \int_y P(y'b(y)|\omega'_y) \propto \int_y P(\omega'_y|y')P(y'|y)b(y). \quad (13)$$

Basic simulations

In this section,[19] we compute posteriors using Equations 3 and 4 to the simple case of a connotative prior that is a single normal distribution with mean μ_y and variance σ_y^2, and a denotative prior that is a binomial distribution $[p_x, 1 - p_x]$ over a discrete set (e.g. "nurse" and "doctor" identities). Starting with equation 4, we have that

$$P'(x) \propto P(x) \int_y P(y) e^{-\frac{1}{2\gamma^2}(y - M(x))^2} dy. \quad (14)$$

[19]Code to generate these plots can be found in the Matlab script `figure2_3_4.m`

https://doi.org/10.1017/S1930297500008652 Published online by Cambridge University Press

Using $P(y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} e^{-(y-\mu_y)^2/\sigma_y^2}$, we complete the squares and integrate out y to obtain

$$P'(x) \propto P(x) \frac{1}{\sqrt{2\pi(\sigma_y^2+\gamma^2)}} e^{-\frac{1}{2(\sigma_y^2+\gamma^2)}(M(x)-\mu_y)^2}. \tag{15}$$

The parameter γ in fact is dependent on x in the general case, and would be written as $\gamma(x)$. We assume this is constant in this paper in order to simplify the presentation, as this may affect the shape of the curves shown in ways that are still to be investigated. This assumption is the same as saying that the variance in sentiments is the same across all identities, for example. While we know this is the case, it is also the case that some of these distributions may be multi-modal or non-normal. We are presenting this here in the simplest way possible to expose the effects of the somatic transform, but show results for $\gamma(x)$ later in this Appendix.

Equation 3, with the prior $P(y)$ as above, is

$$P'(y) \propto \sum_x P(x) e^{-(y-\mu_y)^2/\sigma_y^2} e^{-(y-M(x))^2/\gamma^2} \tag{16}$$

$$= \sum_x \left[\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y-\mu)^2} \right] \left[P(x) \frac{1}{\sqrt{2\pi(\sigma_y^2+\gamma^2)}} e^{-\frac{1}{2(\sigma_y^2+\gamma^2)}(M(x)-\mu_y)^2} \right]. \tag{17}$$

Identifying the second term in [...] with Equation 15 from above,

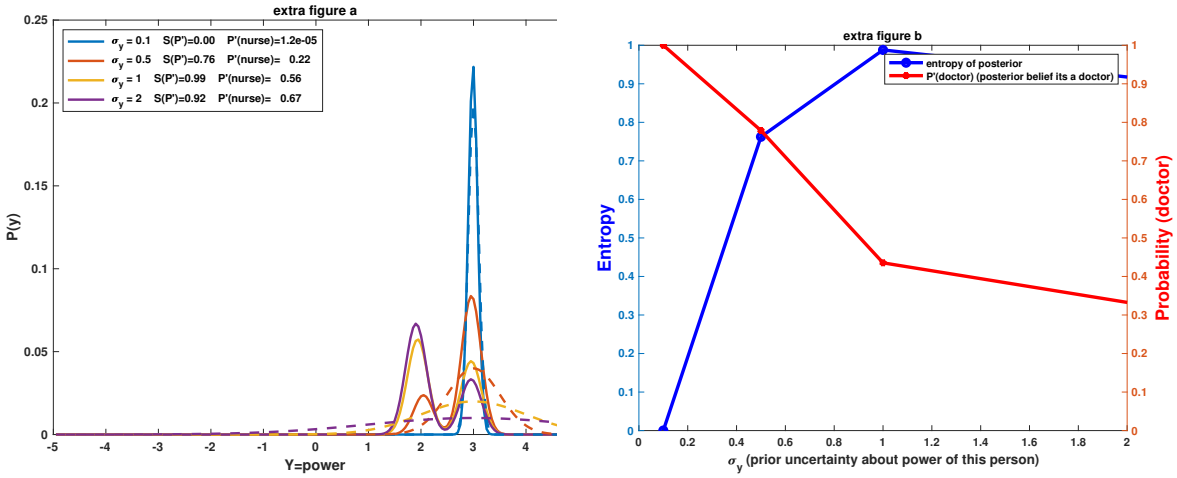
$$P'(y) \propto \sum_x P'(x) \left[\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y-\mu)^2} \right]. \tag{18}$$

where $\sigma = \frac{\sigma_y \gamma}{\sqrt{\sigma_y^2+\gamma^2}}$ and $\mu = \sigma^2 \left(\frac{\mu_y}{\sigma_y^2} + \frac{M(x)}{\gamma^2} \right)$. The term in $[\ldots]$ is the conditional distribution $P'(y|x)$, since $P'(y) = \sum_x P'(x)P'(y|x)$.

Using Equations 15 and 18 with a γ parameter that is not dependent on x, y as the power dimension, and μ_y taken from the USA 2015 dataset, we obtain the figures as in the main text: Figure 2 shows variation with μ_y, Figure 3 shows variation with γ, and Figure 4 shows variation with the prior over x. We further provide two plots showing variation with a changing σ_y in Figures 7 (using $\gamma = 0.2$ and $\mu_y = 3.0$). As the variance in the connotative prior increases, its effect decreases, and the posterior in X resolves to the prior.

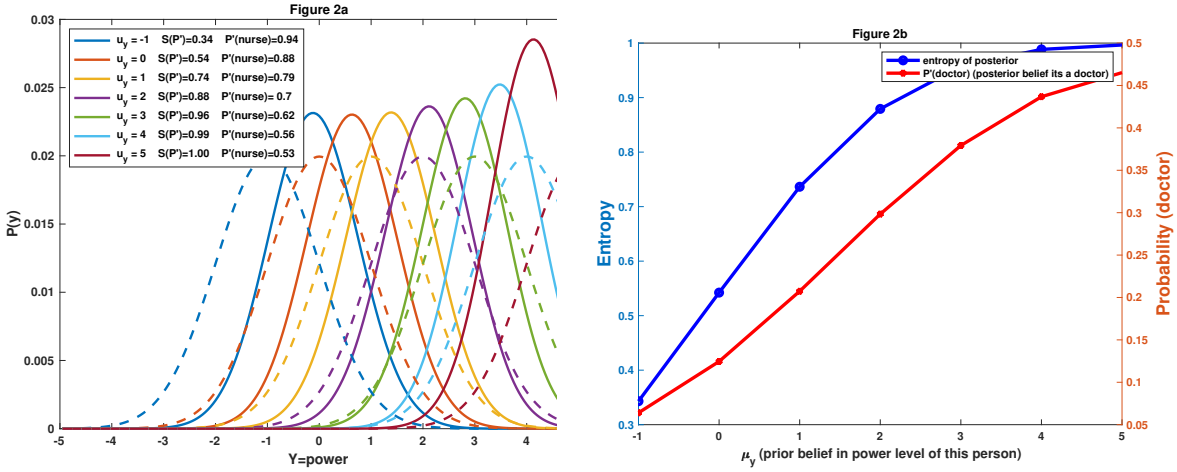
When using $\gamma(x)$ with the actual measurements of variance in the identities from the USA 2015 dataset (variance of 2.5 for *nurse* and 1.4 for *doctor*), we obtain Figure 8 for variation in μ_y, Figure 9 for variation in the prior over x, and Figure 10 for variation in the prior over σ_y. Titles above the figures in what follows give the reference to the equivalent figures in the main text to ease comparisons. The plots are largely the same except for a much smaller variation in Figure 10 due to the larger values of γ. Figures 11, 12, and 13 show the same plots for the evaluation dimension instead of power. These figures are somewhat less interesting as the variations are much smaller due to the similarity in evaluation between nurses and doctors (both in the mean and variance).

https://doi.org/10.1017/S1930297500008652 Published online by Cambridge University Press



(a) (b)

FIGURE 7: Effects of the somatic transform on the posterior marginals in power. (a) Gaussian priors over Y are shown as dashed lines for different values of σ_y. The prior over X is $P(X = nurse) = 0.7$. The posterior over Y is shown as solid lines, while the posterior over X is shown in the legend, with $S(P')$ denoting the entropy of $P'(X)$ and $P'(nurse)$ denoting $P'(X = nurse)$. (b) plot of the overall trends in entropy and posterior probability as a function of μ_y.



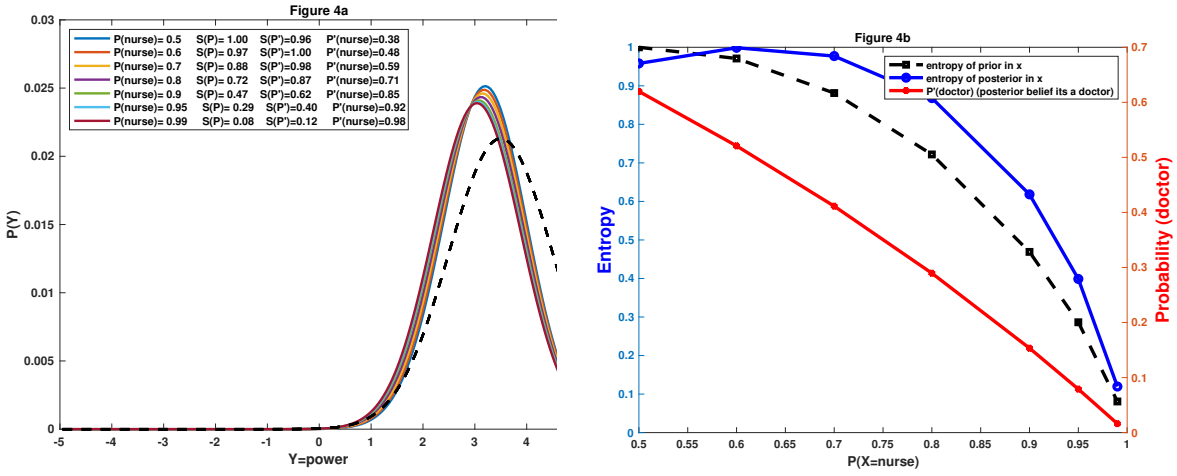
(a) (b)

FIGURE 8: Effects of the somatic transform on the posterior marginals in power for an identity-dependent $\gamma(x)$. (a) Gaussian priors over Y are shown as dashed lines for different values of μ_y. The prior over X is $P(X = nurse) = 0.7$. The posterior over Y is shown as solid lines, while the posterior over X is shown in the legend, with $S(P')$ denoting the entropy of $P'(X)$ and $P'(nurse)$ denoting $P'(X = nurse)$. (b) plot of the overall trends in entropy and posterior probability as a function of μ_y.

Fairness Calculations

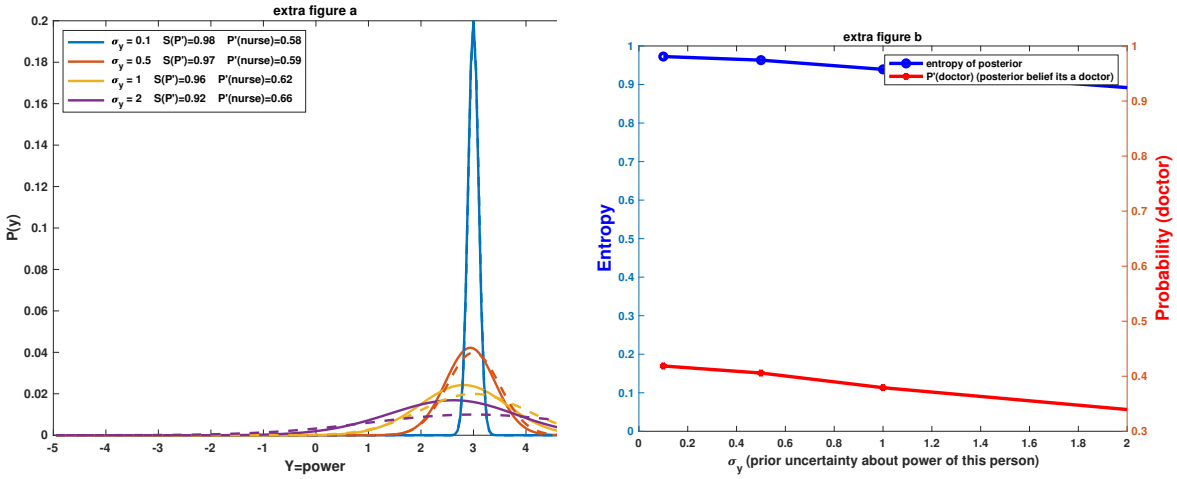
In these simulations, we focus on ACT only in order to find the end-points in Figures 5. The idea is that we compute the emotion that results in each condition (see main text), extract the

https://doi.org/10.1017/S1930297500008652 Published online by Cambridge University Press



(a) (b)

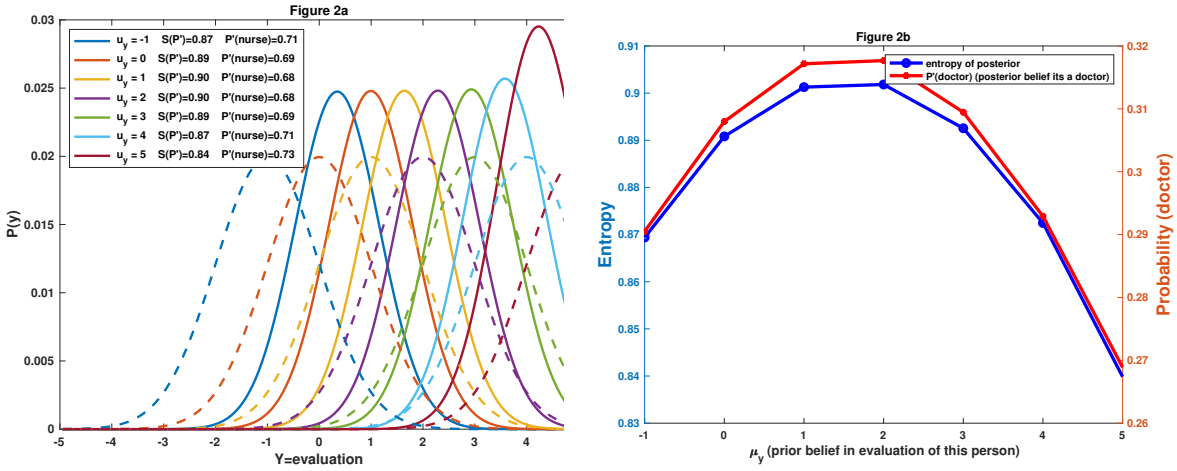
FIGURE 9: Effects of the somatic transform on the posterior marginals in power for an identity-dependent $\gamma(x)$ (a) The red line shows a prior state with a less dispersed $P(X = nurse) = p$ with $p = 0.99$, yielding a posterior for both X and Y that is more in line with the original denotative prior $P(x)$. The blue line ($p = 0.5$) shows how the posterior is biased towards the prior in y (possibly based on stereotypes). The prior in y is shown as a black dashed line (same for all values of p). $S(P)$ and $S(P')$ denote the prior and posterior entropy of $P(X)$, and $P(nurse)$ and $P'(nurse)$ denote the prior and posterior probability of X being *nurse*. (b) plot of the overall trends in entropy and posterior probability as a function of $P(X = nurse)$.



(a) (b)

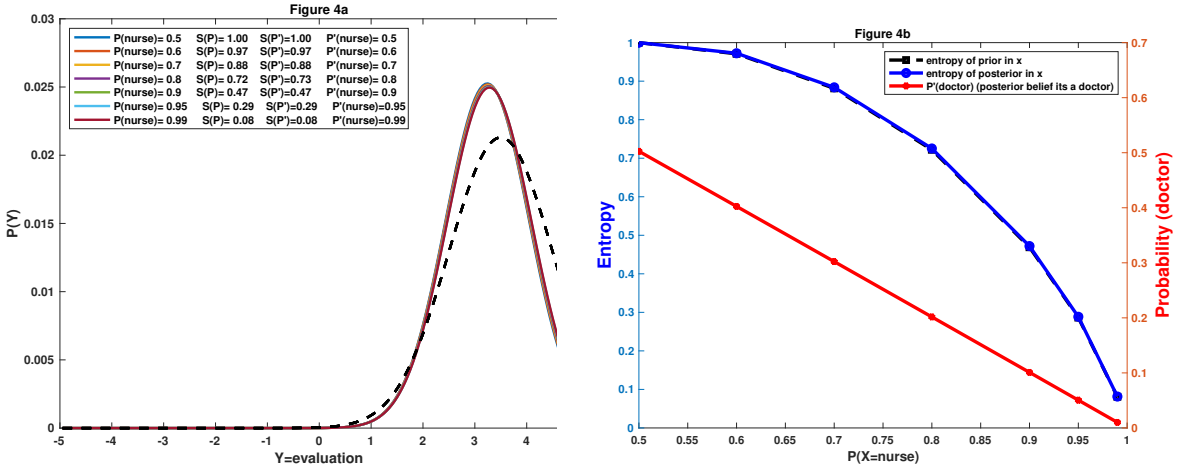
FIGURE 10: Effects of the somatic transform on the posterior marginals in power. (a) Gaussian priors over Y are shown as dashed lines for different values of σ_y. The prior over X is $P(X = nurse) = 0.7$. The posterior over Y is shown as solid lines, while the posterior over X is shown in the legend, with $S(P')$ denoting the entropy of $P'(X)$ and $P'(nurse)$ denoting $P'(X = nurse)$. (b) plot of the overall trends in entropy and posterior probability as a function of μ_y.

https://doi.org/10.1017/S1930297500008652 Published online by Cambridge University Press



(a) (b)

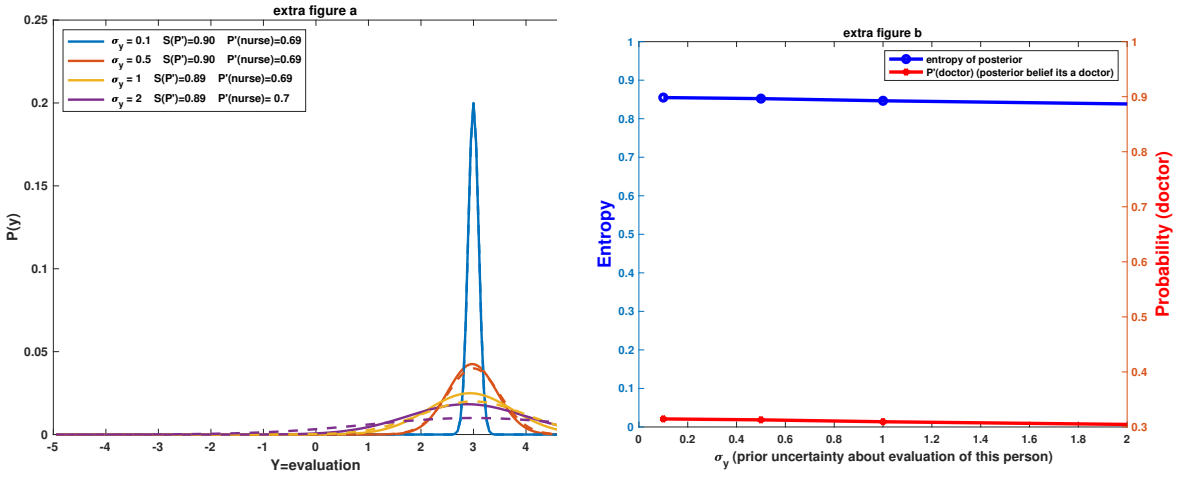
FIGURE 11: Effects of the somatic transform on the posterior marginals in evaluation for an identity-dependent $\gamma(x)$. (a) Gaussian priors over Y are shown as dashed lines for different values of μ_y. The prior over X is $P(X = nurse) = 0.7$. The posterior over Y is shown as solid lines, while the posterior over X is shown in the legend, with $S(P')$ denoting the entropy of $P'(X)$ and $P'(nurse)$ denoting $P'(X = nurse)$. (b) plot of the overall trends in entropy and posterior probability as a function of μ_y.



(a) (b)

FIGURE 12: Effects of the somatic transform on the posterior marginals in evaluation for an identity-dependent $\gamma(x)$ (a) The red line shows a prior state with a less dispersed $P(X = nurse) = p$ with $p = 0.99$, yielding a posterior for both X and Y that is more in line with the original denotative prior $P(x)$. The blue line ($p = 0.5$) shows how the posterior is biased towards the prior in y (possibly based on stereotypes). The prior in y is shown as a black dashed line (same for all values of p). $S(P)$ and $S(P')$ denote the prior and posterior entropy of $P(X)$, and $P(nurse)$ and $P'(nurse)$ denote the prior and posterior probability of X being *nurse*. (b) plot of the overall trends in entropy and posterior as a function of $P(X = nurse)$

https://doi.org/10.1017/S1930297500008652 Published online by Cambridge University Press



(a) (b)

FIGURE 13: Effects of the somatic transform on the posterior marginals in evaluation for an identity-dependent $\gamma(x)$ (a) Gaussian priors over Y are shown as dashed lines for different values of σ_y. The prior over X is $P(X = nurse) = 0.7$. The posterior over Y is shown as solid lines, while the posterior over X is shown in the legend, with $S(P')$ denoting the entropy of $P'(X)$ and $P'(nurse)$ denoting $P'(X = nurse)$. (b) plot of the overall trends in entropy and posterior probability as a function of μ_y.

evaluation e for that emotion, and then compute the average distance to the two emotions of "sad" and "disappointed" using the following rescaling operation:

$$\frac{1}{2}\left[\Big(\big((\text{sad} - e) + 4.3\big) * 6/8.6 + 1\Big) + \Big(\big((\text{disappointed} - e) + 4.3\big) * 6/8.6 + 1\Big)\right]$$

The emotions are computed using the standard equations and the Indiana 2005 dataset. The computer program *Interact* can be used to compute these numbers, for example. We have also made a simple python script to replicate the data in Figure 5, `actsimulator.py`, see http://bayesact.ca for both *Interact* and this script.

Cognitive Dissonance Code

In this section,[20] we compute posteriors using Equations 3 and 4 using priors on y from the USA 2015 dataset for the identity *child*, and from Shank & Lulham (2017) for the two objects: iPhone and blackberry. These simulations are identical to the basic nurse-doctor simulations. However, in this case, the prior over identity is for the *self* (in this case, *child*, and we vary the certainty the has in themselves by changing σ_y. We do this using a prior over X that is skewed towards a "good" item ($[0.8, 0.2]$), which models the *forced-choice paradigm*, shown in Figure 6(a-b), and for an even prior ($[0.5, 0.5]$), which models the *free choice paradigm*, shown in Figure 6(c-d).

[20]Code to generate these plots can be found in the Matlab script `cogdissonance.m`

https://doi.org/10.1017/S1930297500008652 Published online by Cambridge University Press

PCS calculations

In these simulations, we again focus on ACT, and compute the values in Table 1. We compute the evaluation that results in each of the four conditions using an identity of *juror*. Deflections are computed using the standard EPA ratings in the Indiana 2005 dataset and the impression formation equations from the USA 1978 dataset. The computer program *Interact* can be used to compute these numbers, for example. We have also made a simple python script to replicate the data in Figure 5, `actsimulator.py`, see http://bayesact.ca for both *Interact* and this script.

https://doi.org/10.1017/S1930297500008652 Published online by Cambridge University Press