Social motivation in computational neuroscience.
(Or if brains are prediction machines, then the Humean theory of motivation is false)

Matteo Colombo
m.colombo@uvt.nl
Tilburg center for Logic, Ethics and Philosophy of Science
Tilburg University
P.O. Box 90153, 5000 LE Tilburg,
The Netherlands

Abstract Scientific and ordinary understanding of human social behaviour assumes that the Humean theory of motivation is true. The present chapter explores whether and in which sense the Humean theory of motivation may be true in the light of recent empirical and theoretical work in the computational neuroscience of social motivation. It is argued that the Humean theory is false, if an increasingly popular model in computational neuroscience turns out to be correct. According to this model, brains are probabilistic prediction machines, whose function is to minimize the uncertainty about their sensory exchanges with the environment. If brains are these kinds of machines, then we should reconceive the nature of social motivation without appealing to desire. We should rather focus our attention on how social motivation is biased towards reduction of social uncertainty, and on how social norms and other social institutions function as uncertainty minimizing devices.

1 Introduction

Scientific and ordinary understanding of human social behaviour assumes that the Humean theory of motivation is true. According to the Humean theory of motivation, desires and beliefs are "distinct existences"—the existence of one does not imply the existence of the other; and being motivated to act is never merely a matter of having certain beliefs, but also always requires having some desire. In Hume's words: "reason alone can never be a motive to any action of the will… reason is, and ought only to be, the slave of the passions, and can never pretend to any other office than to serve and obey them" (Hume 1978, 413, 415). If this is correct, then social behaviour can never spring from beliefs alone, which represent how things are in the social environment; social behaviour always requires some desire, which specifies how things ought to be in the social environment.

Little attention has been paid to the empirical adequacy of the Humean theory of motivation.[1] In moral psychology and philosophy of action, the case for or against Humeanism has been largely conducted based on arguments that neglect relevant empirical evidence. This way of conducting the debate has been unproductive, and betrays the spirit of Hume's original claim, which is a claim about actual human psychology. In the present chapter, I elucidate the Humean theory of motivation in the light of empirical and theoretical work aimed at uncovering the neurocomputational mechanisms of social norm compliance.

This body of work in computational neuroscience has been helpful in two ways. On the one hand, taking a computational approach forces researchers of social norms to formulate their hypotheses in a precise, quantitative, and self-consistent way. The empirical consequences of hypotheses thus formulated can be fully worked out, and the relationship between theorizing about social norms and empirical testing becomes tight and transparent. On the other hand, neurocomputational modelling allows researchers of social norms to formulate their accounts by integrating data and concepts about the psychology and neuroscience of social behaviour with methods and concepts from the social sciences, and from machine learning and computer science. This blend of ideas and methods has advanced

[1] More attention has received the related, but distinct view of motive *internalism*, according to which moral judgements are intrinsically motivating. Arguments against motive internalism have appealed to evidence about ventromentdial patients and sociopaths, who glide through their social environment with full knowledge of social norms, but without being moved by that knowledge (e.g., Roskies 2003; Kennett & Fine 2008; Colombo 2014a).

understanding of some of the basic computational features of the mechanisms of social norm compliance, promoting integration across disciplines (Wolpert, Doya & Kawato 2003; Behrens, Hunt, & Rushworth 2009; Colombo 2014b).[2]

Drawing on this exciting body of work, I am going to argue that the Humean theory is false, if an increasingly popular model that conceives of brains as kinds of probabilistic prediction machines turns out to be correct (Clark 2013a; Clark 2015a; Hohwy 2013; Friston 2010; Seth 2013). If the Humean theory of motivation is false, then much work directed at understanding the social mind in disciplines like behavioural economics, social psychology, social neuroscience, and moral psychology has been based on a false assumption. We'd better begin scouting the possibility that the social mind is inferential all the way down, and consider how our understanding of the nature of social motivation should change, were the Humean theory of motivation false.

The chapter is in four sections. Section 2 clarifies the commitments of the Humean theory of motivation. Focusing on social norm compliance, Section 3 reviews two types of modelling approaches in computational neuroscience: Reinforcement Learning and Bayesian decision theory. Neurocomputational results within these modelling approaches provide support to the Humean theory in the social domain. Section 4 introduces a third type of approach, *viz*. the "probabilistic predictive processing theory," and shows that it is inconsistent with the Humean theory of motivation. Section 5 suggests how our understanding of social motivation and social norm compliance should change, if the Humean theory of motivation is false and brains are probabilistic prediction machines.

2 The Humean theory of motivation

According to the Humean theory of motivation, belief and desire are distinct kinds of mental states, and belief alone has no motivational force. Believing that the world is thus and so is insufficient for being moved to act. Some desire is always necessary for action; and no other mental states other than a desire and an instrumental belief are necessary for acting, and for rationalizing action (Smith 1994; for a historically informed discussion of Hume's theory and its impact on moral psychology see Radcliffe 2008).[3] A corollary to this claim is that conative states can be changed by the output of some cognitive process only if a conative state features somewhere in this process: "desires can be changed as the conclusion of reasoning only if a desire is among the premises of the reasoning" (Sinhababu 2009, 465). For practical reasoning to have any motivational force, some desire must feature among the premises of the reasoning. If there were no conative state featuring somewhere in the reasoning chain, the conclusion of the reasoning could never affect any of our conative states and could never move us to action.

In order to elucidate the commitments of the Humean theory of motivation, we should clarify the notions of *motivation* and *desire*, and provide some criterion to distinguish *desire* from *belief*.

A *motivation* is any disposition to initiate, direct, or maintain behaviour. *Social motivations* are dispositions to interact with others or in ways that are relevant to others. If you have a motivation to leave a tip at a restaurant after a dinner, then you have a disposition to initiate a certain action directed at certain people in a specific type of context. The motivation to tip at a restaurant is an example of social motivation. If this motivation is not opposed, it will cause you to leave a tip at a restaurant after dinner. If this motivation is opposed by other motivations that dispose you to act differently, or is neutralised by external circumstances, then you may act differently and not leave a tip. Whether you tip or do not tip will impact your waiter, the patron of the restaurant, and probably the other costumers around you.

The causal power of a motivation to control action depends on the strength of the motivation and on its relation to other mental states. You will be moved to act one way or another by your motivations that have greater aggregate strength, and that play an appropriate role in the causal network of mental states leading you to act. If you have a stronger

[2] My focus on neural computation should not suggest that the best (or only) explanation of social norm compliance lies at the level of neural processes. Neurocomputational explanation always spans several levels of analysis. It does not only look at neural processes, it also requires consideration of psychological function, and consideration of the relation between the whole mechanism and its environment across different temporal scales (Casebeer & Churchland 2003).
[3] Belief and desire are the paradigmatic examples of a cognitive state and a conative state, respectively. Unless otherwise noted, I use 'belief' and 'cognitive state,' and 'desire' and 'conative state' interchangeably.

motivation towards leaving a tip at a restaurant after dinner, or if this motivation is appropriately connected to other mental states that lead you to tipping, then you will very probably tip in that restaurant.

Desire is a species of motivation; but not all motivations are desires. Other species of motivation include intentions, interests, appetites, cravings, urges, emotions, and perhaps mental states like the belief that an object or a state of affair is good or the judgement that a certain action is morally right. If desire is a species of motivation, then to have a desire entails having a disposition to act. But to have a desire typically involves more than just being disposed to act in a certain way. It also involves feeling a certain way and thinking in a certain way. If you desire to leave a tip to your waiter after dinner, then you may feel good when you tip him, you may think that tipping is right, or you may have your attention drawn towards information that bears on tipping.

Some desires are intrinsic desires. They are desires that agents have for things for their own sake. Other desires are instrumental. They are desires that agents have for things as a means to achieve another end. For example, people may comply with some social norms for their own sake. They may comply with some social norms "even when there is little prospect for instrumental gain, future reciprocation or enhanced reputation, and when the chance of being detected for failing to comply with the norm is very small" (Sripada & Stich 2007, 285). In these cases, people have an intrinsic desire: they desire to comply with a social norm as an ultimate end, rather than as a means to other ends. Examples include complying with a norm of fairness in one-shot, anonymous dictator games,[4] returning a lost wallet containing a good amount of cash, or punishing norm violators at substantial costs to oneself.

Social desires are desires whose possession or realization is relevant to other agents. In behavioural economics, *social preference* is a related notion that has been extensively investigated. Agents with social preferences do not care exclusively about their own material outcomes; they also care about the material outcomes, well-being, beliefs, and preferences of other agents. A large body of empirical evidence has borne out that people and some other mammals display social preferences in a great many contexts (Fehr 2009).

There are several theories of desire (Schroeder 2014, Sec. 1). One of the most influential is the *reward theory of desire* (Dretske 1988, Ch. 5; Schroeder 2004; Schroeder & Arpaly 2014). According to this account, to desire that something is the case is to use representations of that something to drive reward-based learning. Desire satisfaction would bring about rewards, while desire frustration punishments. Within this account, 'reward' and 'punishment' should be understood as reinforcement learning signals, where rewards strengthen the causal relations between certain mental states and actions, and punishments weaken the causal relations between certain mental states and actions. The dopaminergic system in the midbrain would realize reward-based learning, and its activity would be the common cause of various phenomena traditionally associated with desire, such as action, pleasure, and some aspects of attention and thought.

Drawing on the computational framework of Reinforcement Learning, the reward theory of desire is consistent with the Humean ideas that desire is a distinctive mental state that differs from belief, and that there is a unique causal connection between desire, motivation, and action.

Other accounts of desire are inconsistent with the Humean theory of motivation. For example, according to *desire-as-belief theories of desire*, agents desire something to the extent that they believe it to be good. Accordingly, to the extent you believe it good to comply with a social norm of tipping, you will be motivated to comply with the norm (e.g., Smith 1987; Pettit 1987; Lewis 1988; Broome 1991). This view is inconsistent with the Humean theory because it claims that, for at least some belief, to have a belief is just to have a desire, while the Humean theory holds that belief and desire are "distinct existences," and there is no belief such that to have that belief is to have a desire.

One feature that is commonly used to distinguish desire from belief is their different *direction of fit*. Desire is said to have a direction of fit, and this direction of fit is the opposite to the direction of fit of belief (Anscombe 1957; Humberston 1992). Beliefs are supposed to fit the world. The world is supposed to fit desires.

[4] In the dictator game, a sum of money m is provided to player 1, "the dictator," who determines that x units of the money ($x \leq m$) be offered to player 2. Player 1 retains ($m - x$). Player 2 simply receives x, without having any input into the outcome of the game.

Beliefs aim at truth, whereas desires aim at realization. Beliefs aim at informing us of how things are in the world, whereas desires aim at informing us of how things ought to be in the world. A belief is true if and only if the world is as the belief represents it to be. A desire is realized if and only if the world changes in conformity with the desire. Beliefs must be responsive to evidence that bears on their truth or falsity, they are subject to norms of confirmation and updating. Desires need not be responsive to evidence in the same way as beliefs. If your evidence indicates that the world is not currently as you believe it to be, then you are rationally required to revise your beliefs accordingly. If your evidence indicates that the world is not currently as you desire it to be, then you are not rationally required to change any of your desires.

In summary, for a mental state to count as a desire, it should possess a world-to-mind direction of fit. The world ought to be changed to fit its content. For a mental state to count as a belief, it should have a mind-to-world direction of fit. Its content ought to match the world.

With this distinction in place, we can formulate the central commitments of the Humean theory as follows:

- People possess different kinds of mental states with different directions of fit.

- States and processes with a mind-to-world direction of fit have no motivational force on their own.

- States and processes with a world-to-mind direction of fit are always necessary for action.

I now turn to review two modelling approaches to social norms in computational neuroscience, and show that empirical work within these frameworks support the Humean theory of motivation.

3 Two neurocomputational frameworks for social norms

Social norms constitute a grammar of society (Bicchieri 2006). In nearly all social situations, people's actions are influenced by some social norm.[5] Like a grammar, social norms consist of rules governing social behaviour that need not be written, or legally enforced. Courting and mating, food sharing and dressing, worship and mourning, but also tipping, queueing, playing, and taking vengeance are all paramount cases where some social norm governs people's social behaviour. Like a grammar, social norms specify what is appropriate and what is not appropriate, setting normative boundaries in our social environment. Like a grammar, we often follow social norms without realizing it, or without considering alternative lines of action. In these cases, norm compliance has become second nature to us, helping us to navigate the social environment smoothly and effortlessly. And just like a grammar, social norms are generally not the product of human intentional design. Social norms can emerge (and fade) rapidly and without planning.

Many facts are known about social norms. Social norms are constituted by shared expectations about a certain type of behaviour in a certain context (Bicchieri 2006; Pettit 1990; Sugden 1986). The capacities to learn and follow social norms are developmentally robust and emerge early on in life (Rakoczy & Schmidt 2013). All human societies have social norms, and some non-human animals possess social norms too (Sober & Wilson 1998; Henrich et al. 2004; Colombo 2013). This does not entail that the content of some social norm is invariant across time and space, or that the capacities to learn and to follow norms are underlain by one dedicated mechanism (Casebeer & Churchland 2013).

Rewards, punishments, and emotional sanctions play a major role in social norm compliance. People typically feel anger, contempt, or blame towards norm violators; and norm violators may feel shame and have a desire to hide. Behavioural reactions to norm violators include avoidance, ostracism, gossip, verbal abuse and physical harm (Andreoni et al. 2003; Clutton-Brock & Parker 1995). Nonetheless, norm violations can have political, cultural, and expressive functions. Sometimes, people violate social norms in order to express their attitudes, to raise "consciousness" or to create a public debate about some issue (Sunstein 1996). In these cases, norm violators are

[5] In what follows, the expression 'social norm' is used without distinguishing social norms from other kinds of norms that regulate our social/moral life (for a taxonomy of norms, and for an empirical study of the extent to which different norms are resistant to conformism see Lisciandra et al. 2013).

willing to incur sanctions in order to change the society in which they live, making salient alternative lines of action to which norm compliers may be blind (Hlobil 2015).

Philosophers, social psychologists, anthropologists, and economists have offered different theories of social norms (Bicchieri 2006; Binmore 1994; Elster 1989; Gintis 2010; Lewis 1969; Pettit 1990; Sripada & Stich, 2007; Sugden 1986; Ullmann-Margalit 1977). Most of these treatments consist in what Bicchieri calls "rational reconstructions," where a rational reconstruction "specifies in which sense one may say that norms are rational, or compliance with a norm is rational" (2006, 10-1). Some other accounts provide us with "boxological" models aimed at describing a set of functionally individuated components ("black boxes") of the mechanisms of norm compliance (Sripada & Stich 2007).

While rational reconstructions and boxological models of social norms are valuable for different purposes, neurocomputational approaches have more unificatory power, and they bring more precision to the table when it comes to understanding what goes on in people's heads when they comply with norms (Colombo 2014b).

Two neurocomputational approaches that have been helping researchers to draw precise and unified accounts of social norm compliance are Reinforcement Learning and Bayesian Decision Theory. Both approaches assume strong separation between conative states and cognitive states, where values and probabilities are computed and represented independently. Both approaches assume that conative states (or value signals) are always necessary for action.

Support for the commitments of the Humean theory requires empirical evidence that signals in the brain that vehicle information about rewards and value are distinct from signals in the brain that vehicle information about the probability that a certain state obtains in the environment. It also requires evidence that signals in the brain that vehicle information about rewards and value are always causally involved in the production of action. The good fit demonstrated by Reinforcement Learning and Bayesian models with an impressive body of behavioural and neural data concerning social norm compliance provides this evidence (Glimcher et al. 2009).

3.1 Reinforcement Learning and social norms

Reinforcement Learning (RL) offers models of learning and decision-making in the face of uncertainty and rewards (Sutton & Barto 1998). The type of problem that RL addresses is defined by four basic ingredients $<S, A, T, R>$. First, a set of states $S = \{s_1, \ldots, s_n\}$, where each state is one configuration of the environment. Second, a set of actions $A = \{a_1, \ldots, a_n\}$, which the agent can execute in the environment. Actions can influence the next state of the environment and have different costs and payoffs. Third, a function that governs the transitions between states T: $S \times A \rightarrow [0, 1]$. Given the current state s_t and an action a_t executed by the agent, $T(s_t, a_t, s_{t+1})$ specifies the probability $P(s_{t+1} \mid s_t, a_t)$ of moving to state s_{t+1}. Fourth, a *reward function*: R: $S \times A \rightarrow \mathbb{R}$, which determines the reward r (or payoff) obtained by the agent for executing a certain action in the current state. Specifically, the reward function determines the immediate costs (negative rewards or punishments) and payoffs (positive reward) incurred by performing different actions in the environment. Obtaining a positive reward generates a signal that increases the probability and intensity of the actions that brought about the reward. Instead, obtaining a negative reward (or punishment) decreases the probability and intensity of the actions that brought about the negative reward. Rewards generate approach and consummatory behaviour; punishments generate avoidance.

RL agents' behaviour is captured by a *policy* function π, which defines a mapping from perceived states to actions to be taken when in those states, on the basis of the value of that action in that state. Agents should learn a policy function that makes it most likely that they maximize the total amount of reward they receive in the long run starting from a certain state.

RL modelling has significantly advanced our understanding of the decision-making processes that animals and humans employ when selecting actions in the face of reward and punishment (Niv 2009). Recent work on social norm compliance makes this role particularly salient, and illustrates that results within RL support the Humean theory.

Gu and colleagues (2015) used a RL model to investigate the neurocomputational mechanisms of the capacities to detect norm violations and to change social behaviour accordingly. Their hypothesis was that the insula and the ventromedial prefrontal cortex (vmPFC) play causally necessary, but dissociable roles in the humans' capacities to represent and to adapt to changes in social norms. The focus on these two neural circuits was motivated by previous

work indicating that norm compliance recruits the ventral and dorsal regions of the prefrontal cortex, as well as the cingulate and insular cortex (Anderson et al. 1999; Sanfey 2007; Spitzer et al. 2007).

Gu and colleagues examined the behaviour of four different groups of experimental participants playing the role of responder in an ultimatum game.[6] One group of participants had a focal lesion in the insula; another had vmPFC lesions; a third group included patients with lesions other than vmPFC and insula; a control group presented no brain damage. All participants played several rounds of the ultimatum game. In each round they were offered a split of 20 Chinese Yuan. The offers were drawn from a normal distribution, and presented in a randomized order. Some offers were fair (e.g., 1 Yuan) and others unfair (e.g., 10 Yuan). For each offer, participants could either reject or accept it.

Gu and colleagues found that participants with vmPFC were less sensitive to fairness norms in comparison to the other groups of participants. vmPFC patients were more likely to be motivated to accept unfair offers than the other participants.

To characterise quantitatively the nature of these effects, Gu et al. (2015) used a Rescorla–Wagner RL model (Rescorla & Wagner 1972). The model distinguished between cognitive and conative signals. In particular, it assumed that participants had an internally represented social norm of fairness, and that they could detect norm violations and update the norm representation as follows:

$$[1] \qquad f_i = f_{i-1} + \varepsilon\,(s_i - f_{i-1})$$

where f_i was the social norm of fairness at time step i; s_i was the offer received at time step i; and ε was the norm adaptation rate, which determined the extent to which the social norm could change based on the immediately preceding offer.

The utility $V_i(s_i)$ of an offer for a participant at a certain time was modelled as:

$$[2] \qquad V_i(s_i) = s_i - \alpha \max\{f_i - s_i; 0\}$$

where α determined the extent to which an agent was averse to offers that deviated from the social norm. Based on $V_i(s_i)$, the probability of accepting an offer was modelled according to a softmax function (Sutton & Barto 1998, Ch. 2.3):

$$[3] \qquad P_i(s_i) = \frac{e^{\gamma V_i(s_i)}}{1 + e^{\gamma V_i(s_i)}}$$

where γ is the inverse temperature parameter that controlled how variable were participants' choices ($\gamma \in [0,1]$).

Fitting their model to each participant's choices, Gu and colleagues found that patients with a lesion in the insula had a lower adaptation rate ε and a higher sensitivity α to unequal offers than other participants. These results indicate that both vmPFC and insula play causally necessary, but distinct roles in social norm compliance. Specifically, the insula would play a causally necessary role in learning to adapt to novel social environments, while the vmPFC would play a necessary role for valuing fairness.

Given the good fit of their model with participants' behavioural and neural data, Gu et al.'s (2015) study shows that social norm compliance might be produced by RL mechanisms, which provides support to the Humean theory of motivation.

Consistent with the Humean theory, RL assumes a neat separation between representations of the value of a state and representations of its probability, *viz.*: a separation between value representations and state representations. These

[6] The Ultimatum (or "take-it-or-leave-it") Game is one of the simplest form of bargaining. This two-stage, two-person game is defined as follows. A sum of money m is provided. Player 1 proposes that x units of the money ($x \leq m$) be offered to player 2. Player 1 would retain (m – x). Player 2 responds by either accepting or rejecting the offer x. If player 2 accepts, player 1 is paid (m – x) and player 2 is paid x; if she rejects, each player receives nothing (0, 0). In either case the game is over.

representations play different functional roles in social motivation, and have different directions of fit. Value representations correspond to conative states and are always necessary for action. State representations are representations of the probability of a state, and correspond to cognitive states.

In Gu et al.'s (2015) study, value representations concerned the positive (or negative) charge of different payoff distributions in the ultimatum game. Participants' value representations were captured by the value function $V_i(s_i)$ in [2], which described what the "good" and "bad" offers were for a participant in the long run. [2] also described how the "goodness" of an offer for a participant would change as a function of her previous experience and her attitudes towards unfair offers. The good fit shown by their model indicated that value representations captured by equation [3] always featured in participants' action-selection processes towards realizing high-valued states. The functional role of value representations corresponded to the role of motivating reasons. Value representations caused participants to comply with a social norm, providing a rationalizing explanation of what participants did.

Since RL agents' objective is to maximize the total reward they receive in the long run, the value function $V_i(s_i)$ picked out participants' *goal* in the experiment task designed by Gu and collaborators. If participants had a goal in the task, then they were motivated by a state of mind with the direction of fit of a desire in making their choices. Different decisions in different states had to make the social situation at hand fit participants' goal.

Distinct from value representations were representations of a social norm of fairness f_i, and representations of offers s_i. These were state representations that jointly defined different configurations of the social situation faced by the experimental participants. Representations of norms f_i depicted that a certain social norm was in place with a certain probability. Representations of norms changed as a function of observed offers s_i. Since the offers were explicitly revealed to participants, there was no need to infer s_i from some other observation, or to assign a probability to s_i.

In order to adapt to the social situation, participants had to keep track of the discrepancy between the fair offer f_i and the offer s_i they received, and to update their representations of what counted as a fair offer on the basis of this discrepancy. If the evidence provided by s_i indicated that the social norm f_i was not what participants believed, then they were rationally required to revise their norm representation accordingly. The norm representations f_i of the experimental participants were supposed to fit the changing distribution of offers they received. So, participants' representations of social norms of fairness corresponded to states of mind with the direction of fit of a belief.

3.2 Bayesian decision theory and social norms

Bayesian inference is a type of statistical inference, where data are used to update the probability that a hypothesis is true. Probabilities are used to represent degrees of belief in different hypotheses. At the core of Bayesian inference is a *rule of conditionalization*, which prescribes how to revise degrees of belief in different hypotheses in response to new data.

Consider an agent who is trying to infer the process that generated some data, *d*. Let *H* be a set of (exhaustive and mutually exclusive) hypotheses about this process (i.e., the hypothesis space). For each hypothesis $h \in H$, $P(h)$ is the probability that the agent assigns to *h* being the true generating process, prior to observing the data *d*. $P(h)$ is known as the 'prior' probability. The Bayesian rule of conditionalization prescribes that, after observing data *d*, the agent should update $P(h)$ by replacing it with $P(h \mid d)$ (known as the 'posterior probability'). To execute the rule of conditionalization, the agent multiplies the *prior* $P(h)$ by the *likelihood* $P(d \mid h)$ as stated by Bayes' *theorem*:[7]

$$[4] \qquad P(h|d) = \frac{P(d|h)P(h)}{\sum_{h \in H} P(d|h)P(h)}$$

where $P(d \mid h)$ is the probability of observing *d* if *h* were true (known as 'likelihood'), and the sum in the denominator ensures that the resulting probabilities sum to one. According to [4], the posterior probability of *h* is directly proportional to the product of its prior probability and likelihood, relative to the sum of the products and likelihoods for all alternative hypotheses in the hypothesis space *H*. The rule of conditionalization prescribes that the agent should

[7] Bayes' *theorem* is a provable mathematical statement that expresses the relationship between conditional probabilities and their inverses. Bayes' theorem expressed in odds form is known as Bayes' *rule*. The rule of *conditionalization* is instead a prescriptive norm that dictates how to reallocate probabilities in light of new evidence or data.

adopt the posterior $P(h \mid d)$ as a revised probability assignment for h: the new probability of h should be proportional to its prior probability multiplied by its likelihood.

Bayesian conditionalization alone does not specify how an agent's beliefs should be used to generate a decision or an action. How to use the posterior distribution to generate an action is described by Bayesian decision theory (BDT), and requires the definition of a loss (or utility) function $L(A, H)$. For each action $a \in A$—where A is the space of possible actions or decisions available to the agent—the loss function specifies the relative cost of taking action a for each possible $h \in H$. To choose the best action, the agent calculates the expected loss for each a, which is the loss averaged across the possible h, weighted by the degree of belief in h. The action with the minimum expected loss is the best action that the agent can take given her beliefs.

In the last 20 years or so, also BDT has been playing a tremendous role in advancing understanding of the computational processes underlying a wide variety of cognitive phenomena including social norm compliance (Tenenbaum et al. 2011; on Bayesian unification in cognitive science see Colombo & Hartmann 2015).

Xiang, Lohrenz, and Montague (2013) used a Bayesian model to investigate the neurocomputational mechanisms of social norm compliance. The hypothesis under test was that activity in the anterior insula and in the ventral striatum supports the capacities to detect social norm violations, and to change behaviour accordingly.

Xiang and colleagues examined the behaviour of healthy experimental participants playing the role of responder in the ultimatum game, while undergoing fMRI scanning. All participants played several rounds of the ultimatum game; in each round they were offered a split of 20 American Dollars. The offers were drawn from three different normal distribution with a Low (4 dollars), Medium (8 dollars), or High (12 dollars) mean, and the same standard deviation. To elicit different expectations of fairness, Xiang and colleagues divided their participants into four training groups. Group High-Medium received offers drawn from the distribution with High mean in the first half of their experimental session, and offers from the distribution with Medium mean in the second half. Group Low-Medium received low offers first, and then medium ones. Conversely, group Medium-High and group Medium-Low received medium offers in the beginning, and then high or low offers, respectively.

Participants in different groups exhibited a different pattern of rejection rates for otherwise identical offers. Specifically, in the second half of the experimental session, group Low-Medium rejected medium offers less frequently than group HM. Changes in expectations of fairness led to changes in rejection rates, which correlated with changes in activity in the striatum and anterior insula.

Participants were modelled as Bayesian agents with a prior over offers u, with uncertain variance σ^2 and mean μ. When participants observed an offer x_t at trial t, Bayesian conditionalization was carried out to compute a posterior:

$$[5] \qquad P(u_t|x_t) = \frac{P(x_t|u_t)P(u_{t-1})}{P(x_t)}$$

The mean corresponded to the expected offer at a trial, which was assumed to be the social norm of fairness in the ultimatum game.

To model how participants made decisions, a utility function $U(x_t)$ was defined that quantified the value associated with accepting an offer. According to this function, the value of an offer depended not only on the amount of money offered, but also on the extent to which the offer deviated from the social norm. Utilities were used by a softmax action-selection mechanism to determine choices.

In order to quantify changes in belief about fairness, Xiang and colleagues computed two parameters based on their Bayesian model. Deviations between the expected offer and the offer x_t actually received consisted in a "norm prediction error" parameter:

$$[6] \qquad \delta_t = x_t - \mu_{t-1}$$

A "variance prediction error" parameter captured errors in predictions about the uncertainty (or variance) of the distribution $P(u_t)$. These prediction errors were used in the imaging analysis to uncover the neural signals involved in

social norms violations: the norm prediction errors were positively correlated with activity in the ventral striatum and vmPFC, while the variance prediction correlated with activity in the insula and anterior cingulate cortex.

Xiang and colleagues' (2013) Bayesian model showed a good fit with behavioural and neural data. This provides evidence for the existence of two distinct signals in the brain that vehicle information about social rewards, and about the degree of uncertainty that the social environment is in a certain state. These results support the Humean theory of motivation.

Like in Gu et al.'s (2015) study, Xiang and colleagues (2013) distinguished between norm representations, captured by the mean of the distribution $P(u_t)$, and value representations, captured by $U(x_t)$. Representations of norms consisted of prior beliefs about offers in the game, which would get updated over time as a function of the observed offers x_t via Bayesian conditionalization, as in [5]. Representations of social norms were sensitive to available evidence, and were supposed to fit the changing distribution of offers so that participants had true beliefs about the social situation they were facing.

Posterior distributions were computed independently of expected utilities. Value (or utility) representations picked out what the "good" or "bad" offers were for a participant, consisting of conative states (or preferences) about different offers in game. The values of different offers at a time set participants' goal in the task: maximizing expected utility. Given this goal, participants' representations of value always featured in their decision-making processes, motivating them to realize states with the highest expected utility.

In summary, Gu et al.'s (2015) and Xiang et al.'s (2013) studies, along with an impressive body of literature in computational neuroscience (Glimcher et al. 2009), demonstrate that empirical results within the frameworks of RL and BDT support the tenets the Humean theory of motivation. In particular, these results indicate that social motivation is causally dependent on distinct cognitive and conative states and processes, and that conative states are always involved in social norm compliance.

4 Desiring predictions. The predictive processing theory

In this section, I firstly lay out the basic ideas of the hierarchical, probabilistic, predictive processing (PP) theory of brain and cognition, and then I show that, in its most precise formulation, PP is inconsistent with the Humean theory of motivation.

4.1 PP: Some nuts and bolts

PP claims that brains are homeostatic prediction-testing mechanisms, the central activity of which is to minimize the errors of their predictions about the sensory data they receive from their local environment. Brains would produce perceptual, cognitive, and motor phenomena by minimizing prediction errors courtesy of various monitoring- and manipulation-operations on hierarchical, probabilistic models of the causal structure of the world within a bidirectional cascade of cortical processing (Friston 2009; Friston 2010; Hohwy 2013; Clark 2013a, b; Clark 2015a; Seth 2013). Before explaining how PP conceives of the relation between belief, desire, and action, several concepts deserve clarification (Colombo & Wright 2015).

PP defines *prediction* within the context of probability theory and statistics as the weighted mean of a random variable. The term *prediction error* refers to magnitudes of discrepancies between predictions about the value of a certain variable and its observed value. Within the PP framework, prediction errors quantify the mismatch between expected and observed sensory data. If predictions about sensory data are not met, then prediction errors are generated, which tune brains' probabilistic models of the causal structure of the environment, and reduce the discrepancy between what was expected and what obtained. By minimizing prediction errors, cognitive agents are said to steer clear of *surprising* physiological states that may disrupt their homeostatic properties. *Surprisal*, which is a term from information theory referring to the negative log probability of an outcome, is a function of the sensory data received by an agent, and of a generative model, which is a probabilistic mapping from causes in the environment to observed data.

The most precise formulation of the PP posits that computationally-bounded agents minimize surprisal indirectly, by minimizing *free energy*. Free energy is an information theoretic quantity that bounds the evidence for a model of data. In the context of PP, free energy can be directly evaluated and minimized, and limits (by being greater than) the surprisal on sampling some sensory data given a generative model of these data. By providing a bound on surprisal, minimizing free energy minimizes the probability that agents occupy surprising, maladaptive physiological states (Friston 2009).

The quantities that define the free energy of an agent include: a set B of internal states of the brain b_i, a set D of sensory data d_i, a set A of actions a_i, and a set S of the environmental causes of sensory data s_i. Free energy is a function of sensory data, and brain states that encode a probability distribution of the causes of sensory data. Given a generative model m comprising a likelihood and a prior over sensory data and their causes, the free energy of an agent at a time i is expressed as:

$$[7] \qquad F_i = - \ln P (d_i, s_i \mid m) = - \ln P (d_i \mid s_i, m) - \ln P (s_i \mid m)$$

Minimization of free energy F is said to be driven by changes of brain states b_i and of actions a_i. Both types of changes assume that every self-organizing agent is an embodied model of its sensory exchanges with the environment and environmental niches, and that every agent should maximize the evidence for (or, equivalently, minimizing the uncertainty of) its own model in order to survive and thrive (Friston 2011).

Changes in internal states of the brain can reduce surprisal by minimizing the free energy of current sensory data. This will reduce the divergence between the probabilistic representation of the causes of sensory data encoded by brain states and the true conditional probability distribution of the causes of sensory data. Perception reduces an agent's uncertainty about its sensory exchanges with the environment by making the agent's representations of the environment closer to the truth. If agents' representations of the environment are true, then they are best equipped to avoid surprising states that are potentially noxious.

Changes in action can reduce surprisal by selectively sampling the next sensory datum that, on average, has the smallest free energy. This will ensure that the agent will observe sensory data that are most likely to fit its model of the environment. The basic idea is that action reduces an agent's uncertainty about its exchanges with the environment by selecting courses of action whose sensory consequences are familiar to the agent. If we selectively sample sensory inputs that we most expect, then our expectations become self-fulfilling (Friston et al. 2010).

Thus far, no experimental work has been conducted within the framework of PP that could be compared with Gu et al.'s (2015) or Xiang et al.'s (2013) studies of social norm compliance. Based on a diverse body of evidence from theoretical and simulation studies, PP claims that psychological effects spanning perception, action, cognition, and social behaviour are produced by the same kind of process, *viz*. by the interplay of downward-flowing predictions and upward-flowing sensory signals in the cortical hierarchy in the brain. At each cortical layer, inputs from the previous layer are processed in terms of their degree of deviation from predicted features, and only unexpected features are signaled to the next layer up the hierarchy. Applied iteratively, this processing scheme leads to a two-way direction of processing, where feed-forward connections convey information about the difference between what was expected and what actually obtained, while feedback connections convey predictions from higher processing stages to suppress prediction errors at lower levels (Friston 2010; Hohwy 2013; Clark 2013a).

4.1 PP is inconsistent with the Humean theory of motivation

PP is inconsistent with the Humean theory of motivation for three reasons. First, PP displays action, cognition and perception as unified by a common functional principle, which erodes the distinction between cognitive and conative states and processes. Second, in its free energy formulation, PP "absorbs" utility functions and reward into prior belief, which eliminates conative states as recognizable motivational states. Third and finally, it is not obvious that within PP the direction of fit of cognitive states and processes differ from the direction of fit of states and processes underlying action and motor behaviour.

According to PP, the task of cognitive agents is to represent states of affairs in the world in order to maximize the sensory evidence for their own existence and reduce surprisal (Friston 2011). In order to comply with this task,

perception, cognition, and action all play the same basic functional role: reduction of sensory prediction error. This common functional role furnishes PP with a unifying principle for explaining a great many psychological and neural phenomena. In Friston's words, "if one looks at the brain as implementing this scheme [i.e., free energy minimization], nearly every aspect of its anatomy and physiology starts to make sense" (2009, 293). PP is said to offer "a deeply unified theory of perception, cognition, and action" (Clark 2013a, 186), and even to acquire "maximal explanatory scope" (Hohwy 2013, 242).

If perception, cognition, and action all play the same functional role in producing behaviour, and kinds of mental states and processes are individuated in terms of the functional role they play in producing behaviour, then perception, cognition and action are the same kinds of states and processes. This is exactly what advocates of PP have explicitly asserted. For example, Adams, Shipp, and Friston's (2013, 4) write: "The perceptual and motor systems should not be regarded as separate but instead as a single active inference machine that tries to predict its sensory input in all domains." And further: "The primary motor cortex is no more or less a motor cortical area than striate (visual) cortex. The only difference between the motor cortex and visual cortex is that one predicts retinotopic input while the other predicts proprioceptive input from the motor plant" (Friston, Mattout, & Kilner 2011, 138).

If the processes of perceptual and motor systems, which have been traditionally associated with cognitive and conative functions respectively, do not actually fulfil different functions, but both fulfil the ongoing pursuit of surprisal minimization, then they are identical kinds of processes. This conclusion erodes the distinction between cognitive and conative processes, and is therefore inconsistent with the Humean idea that belief and desire are "distinct existences."

Some may object that this conclusion is too quick. For it does not take account of fine-grained differences between perceptual and motor processes. Clark (2013a, 200) explains that perception and action are "different but complementary means to the reduction of (potentially affect-laden and goal-reflecting) prediction error in our exchanges with the world." On the one hand, perception is said to minimize surprisal by minimizing the free energy of currently observed states. Perception would reduce free energy by optimizing model-based predictions about the causes of sensory data, where predictions are optimal if they are correctly tuned to the actual causes of sensory data. On the other hand, action is said to minimize surprisal by sampling sensory data that have, on average, the smallest free energy. Action would reduce free energy by selecting sensory data that are most likely to fit model-based sensory predictions.

Yet, this fine-grained description will not help to ground a relevant difference between cognitive and conative processes. From the point of view of the scientific taxonomy underlying PP, action and perception differ, but only in ways that are relevant to their function to minimize free energy. Action and perception are *differently the same* from the point of view of the theoretical framework of PP that classifies them. Their functional role is the same. How this functional role gets physically realized differs. If the processes of perceptual and motor systems are differently the same, then these processes should not be said to fulfil different functions.

This conclusion is consistent with the idea that no state or process is motivationally neutral from the perspective of PP. Following this idea, one may suggest that all mental states posited within PP consist of *pushmi-pullyu* signals, which are geared towards the maintenance of agents' homeostatic exchanges with the environment. As explained by Millikan (2004), pushmi-pullyu signals "are undifferentiated between presenting facts and directing activities appropriate to those facts. They represent facts and give directions or represent goals, both at once" (2004, 157). This suggestion is intriguing, but is inconsistent with the Humean tenet that desire and belief are "distinct existences." So, if all states posited within PP consist of pushmi-pullyu representations, or of some other undifferentiated mixture of conative and cognitive states, then PP is inconsistent with the Humean theory.

A second reason suggesting that PP and the Humean theory are inconsistent is the following. PP, at least in its free energy formulation, rejects the core distinction between the probability and the utility (or value) of a state. While this distinction maps onto the distinction between cognitive and conative states, nowhere in the definition of free energy, or in the definitions of action and perception within PP, feature utility, value or reward.

According to PP, agents make decisions by minimizing a free energy bound on the marginal likelihood of observed states given a model of the environment. They do not make decisions by maximizing expected reward. Assuming that hidden states with high probability just are states with high utility, Friston and colleagues show that optimal decision making is tantamount to "fulfilling prior beliefs about exchanges with the world [… while] cost functions are replaced

(or absorbed into) prior beliefs about state transitions" (Friston et al. 2012; Friston et al. 2013; for critical assessments of this proposal see Gershman & Daw 2012, Sec. 5.1, and Colombo & Wright 2015).[8] Free energy is the only quantity that is optimized, and the utility (or value) of a state is not a cause of action, is not a motivating reason. Utility is at best epiphenomenal.

In other words, PP, at least in its free energy formulation, reduces decision-making problems to inference problems. Action aims at producing the most likely sensory data given beliefs about state transitions, instead of bringing about valuable outcomes. This means that within PP "desiring a muffin, for example, is having an expectation of a certain flow of sensory input that centrally involves eating a muffin" (Hohwy 2013, 89). What leads us to eat the muffin is not its associated reward. Rather, it is our expectation about the "familiarity" of the train of sensory inputs associated with eating the muffin. "The 'motivator'—explains Hohwy (Ibid.)—is the urge to minimise prediction error" based on expectations concerning sensory input. But if this is the 'motivator' of action, and *reward* and *value* are eliminated and replaced by prior belief, then PP disqualifies desire as a recognizable mental state that motivates us to act on the basis of the reward (or utility) of a state in the environment.

Even if PP eliminates reward and value, advocates have another option to salvage desire as a recognizable state distinct from belief. They will appeal to the notion of *direction of fit*. Indeed, Hohwy (2013, 89) asserts that "[w]hat makes the desire for a muffin a desire and not a belief is just its direction of fit." Clark (2015, 21) similarly claims that "there remains… an obvious (and important) difference in direction of fit [between perception and action]." The idea is that perception and cognition have a mind-to-world direction of fit because they "match neural hypotheses to sensory inputs," while action and motor control have a world-to-mind direction of fit because they "bring unfolding proprioceptive inputs into line with neural predictions" (Clark 2015, 21; see also Shea 2013).[9]

This way of putting the difference, however, confuses the *direction of fit* of a mental state (or process) with the *direction of causation* of a mental state (or process). Once this confusion is dispelled, it is far from clear that within PP conative processes and cognitive processes have a different direction of fit.

As introduced by Elizabeth Anscombe (1957), the notion of direction of fit is a normative one. Anscombe illustrates the distinction with the different aims of two agents: a man who is guided in his shopping by a shopping list; and a detective who makes a list of the man's shopping items as the man buys them. If we were asked what distinguishes the shopping list from the detective's list, Anscombe explains that "It is precisely this: if the list and the things that the man actually buys do not agree, and if this and this alone constitutes a mistake, then the mistake is not in the list but in the man's performance... whereas if the detective's record and what the man actually buys do not agree, then the mistake is in the record" (Anscombe, 1957, 56). One's beliefs have a world-to-mind direction of fit because they ought to conform to the world. If they do not conform to the world, then the mistake is in the beliefs. One's desires have a mind-to-world direction of fit because the world ought to conform to them. If desires do not conform to the world, the mistake is not in the desires.

This normative notion should not be confused with the notion of direction of causation of a mental state. PP advocates seem to have the latter in mind. They seem to believe that PP makes room for a genuine distinction between cognitive and conative states because in perception the world tends to cause internal changes in one's expectations, whereas in action internal states and processes tend to cause changes in the world that conform to those internal states. However, this is not the notion of direction of fit introduced by Anscombe.

[8] One may object as follows: "PP is not really a competitor of BDT and RL because PP can 'subsume' both, as shown by Friston and colleagues. If the Humean theory is supported by results within BDT and RL, it should also be supported by results within PP. So, PP cannot be inconsistent with the Humean theory." The problem with this objection is that rewards and cost functions are in fact *eliminated* within free energy formulations of PP (e.g., Friston et al 2013; Colombo & Wright 2015). If rewards are eliminated and replaced by prior expectations about occupying different states in the environment, then the Humean theory can be supported by results within BDT and RL—which posit rewards—and be inconsistent with PP—which does not posit rewards.

[9] If the states and processes underlying perception and action have different directions of fit, as asserted by Clark and by Hohwy, then these states cannot be pushmi-pullyu representations as suggested above. Pushmi-pullyu representations have *both* mind-to-world and world-to-mind directions of fit at the same time (Millikan 2004, Ch. 6).

Within PP, agents are regarded as models of their own world (Friston 2010, 127; Friston 2011). As probabilistic models, agents are equipped with a space of hypotheses that are tested against data. Just like scientists' fortunes depend on sound evaluation of scientific models, agents' existence depends on testing hypotheses against sensory data sampled from the world. While agents' hypothesis space is determined by species- and niche-specific evolutionary trajectories (Clark 2013a, 193), the sampling method is dictated by the relative uncertainty of the predictions that the neurally encoded generative model makes about incoming sensory data. Comparing the degree of agreement between sampled sensory data and model-based hypotheses provides evidence about the extent to which one's embodied model fits the world.

Just like in the case of scientific model-based hypothesis testing, agents aim at fitting the world. If model-based predictions do not conform to the world, then the mistake is in the model, which should be updated. If the model is grossly mistaken, and updating unfeasible, then agents' survival is threatened. Instead, to the extent that world and model have a good fit, agents are thereby said to "maximize the evidence for their own existence" (Friston 2010, 136). Both action and perception, therefore, aim at maximize the evidence for our own model of the world. They both aim at fitting the world.

In summary, PP is inconsistent with the Humean theory of motivation. PP erodes the difference between cognitive and conative states by displaying them as unified by a common functional principle. In its free-energy formulation, PP eliminates utility and reward, which are notions distinctive of desire. Third and finally, within PP the direction of fit of cognitive states and processes is not obviously different from the direction of fit of states and processes underlying action and motor behaviour.

5 Social motivation and norm compliance for predictive brains

Now, suppose that the Humean theory of motivation is false, and that PP provides us with a correct picture of how the mind works. What follows about the nature of social norms and social motivation?

It follows that social norms are entropy minimizing devices co-constituted by feedback relationships between minds and world. Human minds would have initial biases towards some "familiar affordances" in their social landscape. These affordances are possibilities for social action provided to agents by the environment—by the patterns, shapes, surfaces, sizes, objects and the other living creatures surrounding the agent. "Familiar" social affordances are features in the environment that are tied to social unfoldings that are likely to maintain one's homeostatic properties within adaptive bounds. Examples of "familiar" social affordances include happy facial expressions, open body postures, and cheerful tones of voice. For example, an extended hand affords a handshake, which is typically tied to co-adaptive social unfoldings like greeting, congratulating, expressing gratitude, or completing an agreement.

By acting on our initial biases towards certain social affordances in different situations, we begin to sample more and more of the sensory space in the social environment. We learn how others generally act in certain contexts, and how they react to our own behaviour. In the same way in which we use models and empirical observation to pick up, refine, and revise hypotheses about the regularities governing the natural world, we would also rely on (embodied) models and empirical observation to uncover the regularities in the social world. We assume certain hypotheses about social unfoldings, we behave as if these hypotheses are true, and the outcomes of our behaviour provide us with evidence that bears on the truth of our hypotheses. As a function of this continuous process of sampling and model testing, our social biases change and our embodied models of the social environment get updated.

As explained by Muldoon, Lisciandra & Hartmann (2014, 4428), "in the social world we react to the real or presumed regularities we identify. However, the social situation is unique in that by looking for regularities, regularities are created… For beings like us, with the psychological tendencies we exhibit in the social world, rule discovery triggers an interest in following the rule. Once this process begins, norms can start to emerge. In this sense, they are created out of nothing, but become real enough once the individuals start to believe they hold true."

When we discover a social regularity, we are thereby attracted to comply with it. Compliance provides us with evidence that confirms our embodied model of the social world, and guarantees that agents engage in "normal," expected

behaviour. When we comply with social norms, we know what to expect from one another, and we are more likely to occupy adaptive sensory states. Along with a language, material symbols, and social institutions like written laws, political structures and religions, social norms help us structure and manage our uncertainty about the outcomes of our interactions in complex social environments.

If PP is true, then social institutions are uncertainty-minimizing scaffolds that constrain and channel people's behaviour cueing expected types of cognitive routines and actions (Schotter 1981; Smith 2007). Thus, social institutions contribute to "normalising" human behaviour making it reliably predictable. In the words of Mary Douglas:

"Institutional structures [can be seen as] forms of informational complexity. Past experience is encapsulated in an institution's rules, so that it acts as a guide to what to expect from the future. The more fully the institutions encode expectations, the more they put uncertainty under control, with the further effect that behavior tends to conform to the institutional matrix […]. They start with rules of thumb, and norms; eventually, they can end by storing all the useful information'' (Douglas 1986, 48).

From the perspective of PP, social norms exist because they help agents minimize the uncertainty of their embodied models of the world. By minimizing uncertainty over their social interactions, agents will tend to become fitting models of the social environment in which they are embedded. The social environment would then become more and more transparent, and social threats easier to avoid (but see Colombo 2014, 74-6, for some qualifications about norm violations, norm ambiguity and normative clash).

In summary, if PP is true, then social motivation is not grounded in desire and value. Social motivation is grounded in minimization of social uncertainty, and in hypothesis testing carried out in different "experiments of social living." The functionality of human social cognition would be best captured by viewing human agents *not* as intuitive lawyers (Haidt 2001), but as intuitive scientists who gather evidence that is most likely to confirm their expectations about the behaviour and mental states of agents co-constituting and continuously re-shaping their social landscapes.

Acknowledgements

I am grateful to Julian Kiverstein, Bryce Huebner, and Chiara Lisciandra for their generous comments on previous versions of this chapter.

References

Adams, R.A., Shipp, S., & Friston, K. J. (2013). Predictions not commands: active inference in the motor system. *Brain Structure and Function*, *218*(3), 611-643.

Anderson, S.W., Bechara, A., Damasio, H., Tranel, D., & Damasio, A. R. (1999). Impairment of social and moral behavior related to early damage in human prefrontal cortex. *Nature*, *2*, 1032-1037.

Andreoni, J., Harbaugh, W., & Vesterlund, L. (2003). The Carrot or the Stick: Rewards, Punishments, and Cooperation. *American Economic Review*, 93, 893-902.

Anscombe, E. (1959). *Intention*, Oxford: Blackwell.

Behrens, T.E., Hunt, L.T., & Rushworth, M.F. (2009). The computation of social behavior. *Science*, 324, 1160–1164.

Bicchieri, C. (2006). *The Grammar of Society: The Nature and Dynamics of Social Norms.* New York: Cambridge University Press.

Binmore, K. (1994). *Game Theory and the Social Contract, Vol. I. Playing Fair.* Cambridge MA: MIT Press.

Broome, J. (1991). Desire, Belief and Expectation. *Mind*, 100, 265–7.

Casebeer, W. D., & Churchland, P. S. (2003). The neural mechanisms of moral cognition: A multiple-aspect approach to moral judgment and decision-making. *Biology and philosophy*, *18*(1), 169-194.

Clark, A. (2015a). *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*. Oxford University Press.

Clark, A. (2015b). Embodied Prediction. In T. Metzinger & J. M. Windt (Eds). *Open MIND:* 7(T). Frankfurt am Main: MIND Group. doi: 10.15502/9783958570115

Clark, A. (2013a). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Science* 36, 131–204

Clark, A. (2013b). The many faces of precision. *Frontiers in Psychology* 4, 270, http://dx.doi.org/10.3389/fpsyg.2013.00270.

Clutton-Brock, T.H., & Parker, G.A. (1995). Punishment in animal societies. *Nature*, 373, 209-216.

Colombo, M. (2014a). Caring, the emotions, and social norm compliance. *Journal of Neuroscience, Psychology, and Economics*, *7*(1), 33-47.

Colombo, M. (2014b). Two neurocomputational building blocks of social norm compliance. *Biology & Philosophy*, *29*(1), 71-88.

Colombo, M. (2013). *Leges sine moribus vanae*: does language make moral thinking possible?. *Biology & Philosophy*, *28*(3), 501-521.

Colombo & Wright (2015). Explanatory Pluralism: An Unrewarding Prediction Error for Free Energy Theorists. http://philsci-archive.pitt.edu/11783/

Colombo, M., & Hartmann, S. (2015). Bayesian cognitive science, unification, and explanation. *The British Journal of Philosophy of Science.* doi: 10.1093/bjps/axv036

Douglas, M. (1986). *How Institutions Think*. New York: Syracuse University Press.

Dretske, F. (1988). *Explaining Behavior: Reasons in a world of causes*, Cambridge, MA: MIT Press.

Elster, J. (1989). Social norms and economic theory. *Journal of Economic Perspectives*, 3(4), 99-117.

Fehr, E. (2009). Social preferences and the brain. In P.W. Glimcher, C. Camerer, R.A. Poldrack, E. Fehr (Eds.). *Neuroeconomics: Decision Making and the Brain*. New York-Amsterdam: Elsevier Academic Press, 215-232.

Friston, K. (2011). Embodied inference: Or I think therefore I am, if I am what I think. In W. Tschacher & C. Bergomi (Eds.). *The implications of embodiment*. Imprint Academic, 89–125.

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

Friston, K. (2009). The free-energy principle: A rough guide to the brain? *Trends in Cognitive Sciences* 13(7):293–301.

Friston, K., Schwartenbeck, P., FitzGerald, T., Moutoussis, M., Behrens, T., & Dolan, R. J. (2013). The anatomy of choice: active inference and agency. *Frontiers in human neuroscience*, *7*.

Friston, K., Mattout, J., & Kilner, J. (2011). Action understanding and active inference. *Biological cybernetics*, *104*(1-2), 137-160.

Friston, K. J., Daunizeau, J., Kilner, J., & Kiebel, S. J. (2010). Action and behavior: a free-energy formulation. *Biological cybernetics*, *102*(3), 227-260.

Gershman, S. J. & Daw, N. D. (2012). Perception, action and utility: the tangled skein. In M. I. Rabinovich, K. Friston, & P. Varona (eds.), *Principles of Brain Dynamics: Global State Interactions* (293–312). Cambridge: MIT Press.

Gintis, H. (2010). Social norms as choreography. *Politics, Philosophy and Economics*, 9(3), 251-264.

Glimcher, P.W., Camerer, C.F., Fehr, E., and R.A. Poldrack (Eds.) (2009). *Neuroeconomics: Decision Making and the Brain*. New York-Amsterdam: Elsevier Academic Press.

Gu, X., Wang, X., Hula, A., Wang, S., Xu, S., Lohrenz, T. M., ... & Montague, P. R. (2015). Necessary, yet dissociable contributions of the insular and ventromedial prefrontal cortices to norm adaptation: computational and lesion evidence in humans. *The Journal of Neuroscience*, *35*(2), 467-473.

Haidt, J. (2001). The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychological review*, *108*(4), 814.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., & Gintis, H. (2004). *Foundations of human sociality: Economic experiments and ethnographic evidence from fifteen small-scale societies*. Oxford: Oxford University Press.

Hlobil, U. (2015). Social norms and unthinkable options. *Synthese*, 1-19. Doi: 10.1007/s11229-015-0863-5

Hohwy, J. (2013). *The predictive mind*. Oxford: Oxford University Press.

Humberstone, I. L. (1992). Direction of fit. *Mind*, 101, 59-83.

Hume, D. (1978). *A Treatise of Human Nature*, second edition, L. A. Selby-Bigge and P. H. Niditch (eds.), Oxford: Clarendon Press.

Kennett, J., & Fine, C. (2008). Internalism and the Evidence from Psychopaths and 'Acquired Sociopaths.' In W. Sinnott-Armstrong (Ed.), *Moral psychology: Vol. 3: The neuroscience of morality*. Cambridge, MA: MIT Press, pp. 173-190.

Lewis, D.K. (1988). Desire as Belief. *Mind*, 97, 323-32.

Lewis, D.K. (1969). *Convention: A Philosophical Study*. Cambridge MA: Harvard University Press.

Lisciandra, C., Postma-Nilsenová, M., & Colombo, M. (2013). Conformorality. A study on group conditioning of normative judgment. *Review of Philosophy and Psychology*, *4*(4), 751-764.

Millikan, R. G. (2004). *Varieties of meaning: the 2002 Jean Nicod lectures*. Cambridge, MA: MIT Press.

Muldoon, R., Lisciandra, C., & Hartmann, S. (2014). Why are there descriptive norms? Because we looked for them. *Synthese*, *191*(18), 4409-4429.

Niv, Y. (2009). Reinforcement learning in the brain. *Journal of Mathematical Psychology*, 53(3):139-154.

Pettit, P. (1990). *Virtus Normativa:* Rational Choice Perspectives, *Ethics* 100, 725-55.

Pettit, P. (1987). Humeans, anti-humeans, and motivation. *Mind*, 96, 530-533.

Radcliffe, E. (2008). The Humean Theory of Motivation and Its Critics. In E. Radcliffe (ed.) *A Companion to Hume*. Oxford, Blackwell Publishing, pp. 477-492.

Rakoczy, H., & Schmidt, M. F. (2013). The early ontogeny of social norms. *Child Development Perspectives*, *7*(1), 17-21.

Rescorla, R.A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In *Classical Conditioning II: Current Research and Theory*, eds A. H. Black and W. F. Prokasy (New York, NY: Appleton Century Crofts, 64–99.

Roskies, A.L. (2003). Are ethical judgments intrinsically motivational? Lessons from acquired sociopathy. *Philosophical Psychology* 16: 51-66.

Sanfey, A. G. (2007). Social decision-making: insights from game theory and neuroscience. *Science*, *318*(5850), 598-602.

Schotter, A. (1981). *The economic theory of social institutions*. Cambridge, MA: Cambridge University Press.

Schroeder, T. (2014). "Desire," *The Stanford Encyclopedia of Philosophy* (Spring 2014 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/spr2014/entries/desire/>.

Schroeder, T. (2004). *Three Faces of Desire*, New York: Oxford University Press.

Schroeder, T. & Arpaly, N. (2014). The Reward Theory of Desire in Moral Psychology. *Moral Psychology and Human Agency: Philosophical essays on the new science of ethics*. Justin D'Arms and Dan Jacobson (eds.) Oxford: Oxford University Press. pp. 186-214.

Seth, A. K. (2013). Interoceptive inference, emotion, and the embodied self. *Trends in cognitive sciences*, *17*(11), 565-573.

Shea, N. (2013). Perception vs. action: The computations may be the same but the direction of fit differs: Commentary on Clark. *Behavioral and Brain Sciences*, *36*(3), 228-229.

Sinhababu, N. (2009). The Humean theory of motivation reformulated and defended. *Philosophical Review*, *118*(4), 465-500.

Smith, M. (1994). *The Moral Problem*. Oxford: Blackwell Press.

Smith , M. (1987). The Humean theory of motivation . *Mind* , 96 , 36 – 61 .

Smith, V. (2007). *Rationality in economics: constructivist and ecological forms*. New York: Cambridge University Press.

Sober, E., & Wilson, D. S. (1999). *Unto others: The evolution and psychology of unselfish behavior*. Cambridge, MA: Harvard University Press.

Spitzer, M., Fischbacher, U., Herrnberger, B., Grön, G, & Fehr, E. (2007). The Neural Signature of Social Norm Compliance. *Neuron*, 56, 185-196.

Sripada, C., & Stich, S. (2007). A framework for the psychology of norms. In P. Carruthers, S. Laurence, and S. Stich (Eds.), *The innate mind: culture and cognition*. Oxford: Oxford University Press, 280–301.

Sugden R. (1986) *The Economics of Rights, Cooperation and Welfare*. Oxford: Blackwell.

Sunstein, C. R. (1996). Social norms and social roles. *Columbia law review*, 903-968.

Sutton, R.S., & Barto, A.G. (1998). *Reinforcement learning: an introduction*. Cambridge, MA: MIT Press.

Tenenbaum, J.B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, *331*(6022), 1279-1285.

Ullmann-Margalit, E. (1977). *The Emergence of Norms.* Oxford: Oxford University Press.

Wolpert, D.M., Doya, K., & Kawato, M. (2003). A unifying computational framework for motor control and social interaction. *Philosophical Transactions of the Royal Society of London B Biological Sciences*, 358, 593–602.

Xiang, T., Lohrenz, T., & Montague, P. R. (2013). Computational substrates of norms and their violations during social exchange. *Journal of Neuroscience*, 33, 1099-1108.