Review

The sense of should: A biologically-based framework for modeling social pressure

Jordan E. Theriault [a,*], Liane Young [b], Lisa Feldman Barrett [a,c,d]

[a] *Department of Psychology, Northeastern University, Boston, MA, USA*
[b] *Department of Psychology, Boston College, Chestnut Hill, MA, USA*
[c] *Psychiatric Neuroimaging Division, Department of Psychiatry, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA*
[d] *Athinoula A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Boston, MA, USA*

Abstract

What is social pressure, and how could it be adaptive to conform to others' expectations? Existing accounts highlight the importance of reputation and social sanctions. Yet, conformist behavior is multiply determined: sometimes, a person desires social regard, but at other times she feels obligated to behave a certain way, regardless of any reputational benefit—i.e. she feels a *sense of should*. We develop a formal model of this *sense of should*, beginning from a minimal set of biological premises: that the brain is predictive, that prediction error has a metabolic cost, and that metabolic costs are prospectively avoided. It follows that unpredictable environments impose metabolic costs, and in social environments these costs can be reduced by conforming to others' expectations. We elaborate on a *sense of should*'s benefits and subjective experience, its likely developmental trajectory, and its relation to embodied mental inference. From this individualistic metabolic strategy, the emergent dynamics unify social phenomenon ranging from status quo biases, to communication and motivated cognition. We offer new solutions to long-studied problems (e.g. altruistic behavior), and show how compliance with arbitrary social practices is compelled without explicit sanctions. Social pressure may provide a foundation in individuals on which societies can be built.

Nature, when she formed man for society, endowed him with an original desire to please, and an original aversion to offend his brethren. She taught him to feel pleasure in their favourable, and pain in their unfavourable regard.
. . . .

But this desire for the approbation, and this aversion to the disapprobation of his brethren, would not alone have rendered him fit for that society for which he was made. Nature, accordingly, has endowed him, not only with a

---

* Corresponding author at: Department of Psychology, Northeastern University, 140 Nightingale Hall, Boston, MA, USA.
E-mail address: jordan_theriault@northeastern.edu (J.E. Theriault).

desire of being approved of, but with a desire of being what ought to be approved of . . . . The first desire could only have made him wish to appear to be fit for society. The second was necessary in order to render him anxious to be really fit. Adam Smith [268, III, 2.6–2.7]

1. Introduction

How does social pressure work? And what benefit does an individual gain by conforming to others' expectations (e.g. expectations to help others, [251]; expectations to hurt others, [89]; or even innocuous expectations, like suppressing a cough in a quiet hallway)? Conformity in the face of social pressure is a well-known behavioral phenomenon [5,6,133,201,205] and is multiply determined [27,74,75,251]. For example, if you and a group of others were asked a question, and if all other group members gave a unanimous response [5,6], then if you copied the group's answer at least two sources of influence might have motivated your behavior: you might have copied them because you assumed they were knowledgeable (i.e. you experienced *informational influence*), or you may have copied them despite knowing they were incorrect (i.e. you experienced *normative* influence; [74,75,285]). In this paper, our aim is to elaborate on how normative influence motivates behavior. Typically, it is assumed that normative influence motivates individuals through actual or anticipated social rewards and punishment (e.g. reputation, social approval; [46,55,167,216,85,285]; but see [133]). That is, one individual conforms to another's expectation (or to expectations shared collectively; i.e. norms; [32,143]) to "gain or maintain acceptance" [167, p. 411], to avoid "social sanctions" ([46, p. 1015]; see also, [251, p. 225]), to achieve "social success" [217, p. 556], or to "signal belongingness to a group" [285, p 580].

But this explanation cannot be complete. For one, non-conformists are frequently popular [205, Chapter 4], which implies that individuals sometimes gain acceptance or achieve social success by violating expectations and norms. But more importantly, just as conformist behavior is multiply determined (by informational and normative influence; [74]), normatively motivated behavior is multiply determined too. As Adam Smith observed (among others; e.g. [7, Chapter 12], [27,75,133,221,251,286]), a person is motivated *both* by a desire for social regard, and by a sense that she *should* behave a certain way. If a person were only motivated by reputation (i.e. *reputation-seeking*), then she would only be motivated to *appear* norm compliant [268, III, 2.7]. People can, indeed, be motivated by reputation-seeking (e.g. when they explicitly select behaviors that will make others like them). However, in this paper we focus on Smith's second motivation—the motivation to match one's behavior (i.e. conform) to individual others' expectations or to the norms of a culture, without expecting or aiming to bring about a social (e.g. reputational) or non-social (e.g. money, food) reward. We call this felt obligation to conform to others' expectations a *sense of should*.

Adam Smith highlighted that reputation-seeking and obligation are separable motives; however, he and others (e.g. [286]) do not distinguish between moral and non-moral (i.e. social) obligations (Fig. 1). For present purposes, we distinguish these influences on the basis of whether others' expectations motivate behavior. A *sense of should* refers to a felt social obligation to conform to others' expectations. By contrast, following [251], we use moral obligation to refer to cases where an action is motivated by an internalized personal value, even when the action would violate others' expectations—for example, a moral obligation "to tell the truth even if it is painful" [7, p. 356] motivates an individual to violate others' expectations, opposing a *sense of should*. In this paper, we are centrally concerned with how others' expectations motivate behaviors via a *sense of should*, independent of reputation-seeking or internalized personal values.

Social pressure and its subjective experience (a *sense of should*), then, describes something much more common than moral obligation. A *sense of should* may motivate you to observe arbitrary, typically unenforced social customs (e.g. wearing nail polish if female), to tolerate physical discomfort in social settings (e.g. waiting to go to the restroom during a lecture), and to follow others' commands (e.g. passing the salt when asked). This motivation to conform to others' expectations may be the social scaffolding that makes society possible ([94,133,208,255,257]; also see [80]); yet, as to why individuals conform to expectations, feel social obligations, or accept social institutions: "there does not seem to be any general answer" [257, p. 108]. Our aim in this paper is to address this question from evolutionary and biological principles, beginning with empirical research in neuroscience and neuronal metabolism, building to a formal account of a *sense of should*'s function and proximate mechanisms, and ending with an outline of how this individual motivation might emergently produce social phenomenon ranging from communication, to status quo biases, culture, and motivated cognition.

Category of Influence		Do you copy others' behaviors?	Are you motivated by others' expectations?	Do you feel obligated to perform/not perform a behavior?
(Normative Influence)	Informative Influence	Yes	No	No
	Reputation-seeking	Not necessarily	Yes	No
	Sense of Should (i.e. Social Obligation)	Not necessarily	Yes	Yes
	Moral Obligation	No	No	Yes

Fig. 1. Diagram of relevant influences on behavior. *Informative influence* refers to the "wisdom of the crowd", where an individual copies others' behavior because she assumes they are knowledgeable [74]. *Normative influence* motivates compliance with others' expectation, but it does not necessarily motivate copying their behaviors—i.e. an individual may copy others' behavior to fit in socially [5,74,167], or she may help a victim because others expect her to [252,253]. Within normative influence, we distinguish *reputation-seeking*, where an individual explicitly aims to receive praise or avoid blame, from a *sense of should*, where an individual feels obligated to conform to others' expectations. *Moral obligation* refers to cases where an individual feels obligated to perform a behavior, but is motivated by something besides others' expectations (e.g. personal values; [251]). Note that a behavior (or category of behaviors) may be typically called "moral" (e.g. sharing) but the behavior could be motivated by any of these influences. This list of influences is also not exhaustive.

1.1. A biologically-based sense of should

In our framework, a *sense of should* refers to a felt obligation to conform to others' expectations. We will suggest that a *sense of should* is learned, and is experienced as an anticipatory anxiety toward violating others' expectations (see also, [75,221]). We hypothesize that this anticipatory anxiety stems from the unpredictable social environment (and its affective consequences) that an expectation-violating behavior is anticipated to create. That is, when you violate others' predictions about your behavior, we hypothesize that their behavior becomes (or is anticipated to become) more difficult to predict for you.

In this paper, we ground our account of a *sense of should* in a biologically plausible evolutionary context. Prior research in evolutionary psychology and behavioral economics has also acknowledged motives beyond reputation-seeking, noting that behaviors can be motivated by "irrational" (typically emotional) sources, which are experienced as distinct from rational, self-interested motivations. For example, responding with anger might be a more effective deterrent against cheaters, compared to dispassionately deciding whether to retaliate [96]. Or, self-deception might insulate your consciousness from the true motives driving your behavior, helping you more easily deceive others [294, 289]. Or, emotions that lead you to cooperate without considering costs may signal to others that you are a trustworthy partner, securing future reciprocal exchanges [151]. As evolutionary models, these all provide detailed accounts of the ultimate benefits; however, they provide sparse accounts of the proximate mechanisms. For example, it is taken as a sufficient explanation that "negative emotions" [82], or "moral outrage" [162] motivate prosocial punishment. These appeals to emotion are a route to a black box—they offer no further explanation of the proximate mechanism, only a description and a label. That is, given that decades of research has failed to identify any consistent neural architecture implementing discrete emotional experiences [22,51,137,301], there is no clear path to pursue proximate accounts of "negative emotions" or "moral outrage" to the biological level on which natural selection operates.

By contrast, modern accounts of emotion have suggested that emotions derive from a combination of bodily (interoceptive) sensation (signals from the body to the brain indicating, for example: heart-rate, respiration, metabolic and immunological functioning; [26,58,261]) and a brain capable of categorizing patterns of sensory experience [19, 21–24,240,241]. By leveraging these advances in the study of the brain and emotional experience, we can provide a full evolutionary account, showing how, at an ultimate level, individual fitness is promoted by conforming to others' expectations, and how, at a proximate level, this *sense of should* works. Importantly, this evolutionary account does not depend on the plausibility of discrete, functionally specific adaptations (i.e. modules; [56]). Instead, we suggest that a *sense of should* is an emergent phenomenon, and could arise from domain-general developments (e.g. in the capacity for inference and memory, in combination with a social context). This domain-general account also raises the possibility that prosocial behavior in humans is *not necessarily* made adaptive by the long-term benefits of reciprocal

altruism [141,288]. Rather, behaving as others expect may be adaptive as a simple consequence of the immediate biological benefits of a predictable social environment.

To explain a *sense of should*, we will situate our approach in the context of a biological common denominator: energy consumption (i.e. metabolics). Humans, like all organisms, are *resource rational* [136,188]: they optimize their use of critical resources, which, for living creatures, are metabolic. At a psychological level of analysis, behavior can be understood as driven by distinct motives—e.g. "self-interest" (such as reputation-seeking) vs. a *sense of should*. However, we suggest that social behavior may be more systematically understood by beginning at a deeper level of analysis, a level where both "self-interest" and a *sense of should* act as strategies for satisfying the energetic needs of the organism.

Many researchers are accustomed to considering evolutionary fitness only in terms of reproductive success (e.g. [67]; but see, [302]); however, "at its biological core, life is a game of turning energy into offspring" [222, p. 170], meaning that for all organisms the management of metabolic resources is central—reproduction is one metabolic investment among many.[1] We will suggest that a *sense of should*, like "self-interested" motivation in the traditional sense, is adaptive because it allows humans to manage the metabolic demands imposed by their social environment. Both motives are self-interested in an ultimate sense, and provide complementary routes to the same adaptive end.

1.2. Outline

In this paper, we use a biological framework to develop a mechanistic account of the *sense of should*. We address why people are motivated to conform to others' expectations, and make our logic clear in a formal mathematical model. We begin by outlining the biological foundations of our approach (section 2), applying key insights from cybernetics [54,235] and information theory [266] to characterize the brain as a predictive, metabolically-dependent, model-based regulator of its body in the world. For humans, this world is largely social, and at the core of our approach is the hypothesis that individuals make this social environment more predictable by inferring others' expectations and conforming to them. By conforming, an individual can regulate others' behavior, the rate of her own learning, and the metabolic costs imposed by her social environment. We formalize the individual adaptive advantages of this strategy (section 3), then elaborate on the proximate psychological experience of a *sense of should*, the precursors for its development, its relationship to mental inference, and what is unique about this indirect form of influence. Finally, we explore the potential for our framework to unify disparate evolutionary, anthropological, and psychological phenomena (section 4), including status quo biases, communication, game-theoretic explanations of behavior, and the inheritance of culture and social norms. Taken together, this paper aims to begin from biological principles, and end with a unified framework to describe socially motivated behavior.

2. Biological foundations for a *sense of should*

The biological foundations for a *sense of should* involve a general account of what a brain is for and how it regulates the body's interactions with the world. In this section, we review established work in neuroscience and introduce key concepts related to brain energetics—the metabolic processes that power neural activity. We show that organisms promote their own survival by using a predictive, regulatory model (i.e. a brain) to ensure that interactions with their environment are metabolically efficient. A logical consequence is that unpredictable environments are metabolically costly. With this foundation in place, we suggest that the human brain also regulates the metabolic costs of its *social* environment, via a *sense of should*.

To some readers, it may seem unintuitive, or even reductive to ground motivation in metabolism (but see [45]). However, it must be remembered that Western, Educated, Industrialized, Rich, and Democratic people (i.e. W.E.I.R.D.; [146]) are spoiled for resources in a way that is unprecedented among past and present human societies (let alone the animal kingdom). We (or, we who are economically secure professors and professionals) are cushioned by grocery stores, houses, and a culture that sustains them, meaning that calculations balancing fighting,

---

[1] Organisms also adopt strategies that differently weight reproduction and survival [281]: viruses reproduce quickly, and adapt to their environment on a generational timescale, whereas humans and other brained organisms reproduce slowly, but adapt to their environment within one lifetime. The slower reproductive strategy of animals requires that energy consumption be regulated to promote long-term survival (which, in turn, provides more potential opportunities for mating and reproduction).

fleeing, and feeding are not currently experienced as pressing concerns. These calculations may not be salient to us, but they are central to the evolutionary history of all organisms, and within behavioral ecology a gain or loss in metabolic efficiency can determine whether an individual, or even a species, survives [37,175]. W.E.I.R.D. culture may buffer many metabolic concerns, but we suggest that these concerns nonetheless shaped our evolutionary history, forming the psychological processes that allowed society to emerge. If we want to understand how society is maintained—and how a life of metabolic leisure is supported—then we must begin from these biological principles.

2.1. A brain regulates a body in its environment

As an organ common to humans, flies, rats, and worms, a brain has a common purpose, shared across species: to regulate a body in its interaction with the environment [22,26,203,236,281]. Fundamentally, the brain's job "reduces to regulating the internal milieu and helping the organism survive and reproduce" [281, p. 11], a conjecture supported by evidence from neuroanatomy [41,174], and from neural physiology and electric signal processing [281]. Of course, regulation varies in its particulars—the innards and environs of worms and humans pose drastically different regulatory challenges—but the core regulatory role of the brain remains unchanged. On this account, sensation and cognition are functionally in the service of this regulation—they are the means to an end: what you see, feel, think, and so on, is all in the service of the brain regulating its body's interactions with the world.

At the core of regulation lies the management of metabolic processes. To survive, grow, thrive, and ultimately reproduce, an organism requires a near continuous intake of energetic resources, such as glucose, water, oxygen, and electrolytes—it must be watered and fed. Resources maintain the body and fuel physical movements, movements that can acquire more resources or protect against potential threats. All actions have some metabolic cost, but to acquire more resources organisms must forage or hunt. What this means is that survival is not a matter of minimizing metabolic expenditures—instead, organisms must be efficient: they must invest energy to provide the largest metabolic return.

The brain itself is a significant energy investment. In rats, it accounts for ~5% of energy consumption; in chimpanzees ~9%; and in humans, ~20% ([50,152]; and this percentage is even higher in children, see [131,169]). Cognitive functions, such as learning, are metabolic investments also: they require energy in the form of glucose and glycogen (e.g. [148]), which are metabolized to produce neurotransmitters—e.g. glutamate [113]—and ATP molecules [199], the foundational energy source for the brain. In times of scarcity, learning may be a poor investment and may be limited to features that promote survival in the short term. But in times of abundance, an organism can promote its own survival by learning and exploring the environment [39], finding safer or more metabolically efficient ways to exploit it [53]. This interplay between conservation of energy during scarcity, and investment during abundance, is critical to keep in mind. In explaining a *sense of should*, we will be largely focused on methods of conservation; yet, exploration (including sometimes violating expectations and norms) will serve a critical role in learning (see Section 2.3.1 Constructing and Coasting).

This idea, that metabolic resources should be spent frugally and invested wisely, dates at least as far back as Darwin, who observed that "natural selection is continually trying to economize in every part of the organization" and that "it will profit the individual not to have its nutriment wasted on building up [a] useless structure" [64, p. 137]. If a costly biological structure provides no return (i.e. it does not promote survival or reproduction) then evolution should select against it. This logic can be extended to behavior and cognition, implying that an organism's cognition should only be as complex as is necessary for it to survive in its ecological niche [119–121]. This observation foreshadows our hypothesis: the human ecological niche is *social*; and therefore, the social environment profoundly affects which behaviors and cognitions are energetically optimal.

As a good regulator, the brain facilitates survival by modeling the environment, and at the same time it must not spend more energetic resources than necessary. In the next section, we discuss how a brain promotes survival by acting as an internal model of its body and environment. In section 2.3, we discuss how a brain uses efficient, predictive processing schemes to minimize the metabolic costs of neuronal signaling. Then, in section 3, we return to the social world, demonstrating how a *sense of should* motivates humans to manage the metabolic costs imposed by other people.

2.2. Allostatic regulation: the brain is a predictive model

A brain regulates its body, and in doing so it should avoid costly mistakes. For example, when threatened, a coordinated suite of fight-or-flight responses are deployed in a context-sensitive way (e.g. raising blood pressure; redirecting bloodflow from kidneys, skin, and the gut to muscles; increasing synthesis of oxidative enzymes and decreasing production of immune system cells; [193,280,298]; see also, [25]). Critically, an organism must implement these bodily changes *before* a predator's teeth close around its neck—it must respond to the anticipated harm, not the harm itself. Likewise, even getting up from a chair requires a redistribution of blood pressure before you stand (i.e. a slight rocking head motion induces vestibular activity, raising sympathetic nervous activity before standing; [98]), or else the error, "postural hypotension", will cause fainting and perhaps a sprain or a broken bone [278]. Mistakes can be dangerous, even deadly, and a good regulator must avoid serious errors.

A core principle of cybernetics makes clear how this challenge is met: "Every good regulator of a system must be a model of that system" [54]. Your body, in its interaction with the environment, is the system in question, and your brain is the internal model of that system—i.e. its regulator [22,23,54,236,262]. The best models learn: they modify themselves when mistakes occur so that they can predict better in the future [235]. To regulate efficiently, then, the brain must regulate predictively—it must anticipate outcomes and direct behavior accordingly.

This predictive regulation is called *allostasis* [249,278–280], where a brain anticipates the needs of the body and attempts to satisfy those needs *before* they arise, minimizing costly errors. For instance, organisms should be motivated to forage *before* vital metabolic parameters (e.g. glucose, water) run out of safe bounds [278]. Allostatic regulation stands in contrast to the more familiar *homeostatic* regulation, where parameters are kept stable around a set-point, e.g. as in a thermostat, which cools the room when it gets too hot and warms it when it gets too cold. For any living organism homeostatic regulation is risky: it only occurs in reaction to events, meaning that it *must wait for errors to occur* [54]. With a brain (i.e. a model of the system), such errors can be avoided [22,54,262]: by modeling the system, organisms can adapt to environmental perturbations *before* they occur. Allostasis then, is powerful because it is predictive—a model anticipates challenges and prepares the organism to meet them.

Evidence for allostasis is hidden in plain sight, just below the surface of familiar experimental paradigms. For instance, when shocks are delivered to rats, stress-induced physiological damage is minimized when a cue makes shocks predictable. Compared to unsignaled shocks, signaled shocks halved the size and quantity of resultant ulcers, even when the signaled shocks could not be escaped or avoided [299]. Further, some of the most compelling evidence for anticipatory regulation comes from Pavlov. Pavlov's classic experiments—where dogs first salivate to the food stimulus, and later to the conditioned stimulus of the dinner bell—are commonly taken as evidence for a reactive, stimulus–response driven psychology. But Pavlov's Nobel prize was awarded for his work in physiology, where he demonstrated that both before and during feeding, the dog's saliva and stomach acid is prepared with the appropriate mix of secretions to facilitate digestion [112,218,281]. For fats, lipase is prepared in the mouth and bile in the stomach. For bread, starch-converting amylase is secreted with saliva. For meats, acid and protease accumulates in the stomach. In each case, the brain predictively coordinates a suite of bodily responses: when food enters the stomach, it meets an environment already prepared to metabolize it.

Allostasis implies that all organisms use a model to guide behavior. This conclusion may appear to conflict with recent work in reinforcement learning, which suggests that organisms switch between "model-based" (i.e. goal-directed) and "model-free" (i.e. habitual) modes of learning ([60,62,65,66,204]; but see [101]). Specifically, model-free learning does not create a plan to reach a goal (e.g. the "cheese" in a maze); instead, it reinforces discrete actions (e.g. move left, move right) through a repeated process of trial-and-error. But, as said above, trial-and-error strategies are inherently dangerous. Organisms will sometimes make mistakes, and when these mistakes occur, organisms should learn from them; however, organisms should never completely abandon the internal model into which they have continually invested metabolic resources, reverting to a pure trial-and-error strategy. (Of course, it would be plausible to consider model-free and model-based strategies along a spectrum, from short-term to long-term model-based strategies, in which case our point is simply that organisms never completely move to the model-free pole.) In computational simulations model-free strategies can learn across millions of trials, but for a living organism each mistake could be fatal, bringing learning to a premature end (e.g. [308]).

The appeal of model-free learning typically stems from an assumption that it is computationally cheap compared to a model-based strategy. For example, it is sometimes assumed that a model-based strategy involves activating a brain-region (e.g. prefrontal cortex) and engaging in an expensive search through a goal-directed decision tree

[66,238]. But this perspective misunderstands how and when living organisms pay down the cost of their internal model. Learning consumes metabolic resources [113] to construct and modify a neural architecture. But the cost of creating this neural architecture is distributed over the course of a lifetime [131,169,202,203]—you have been investing in an internal model of your ecological niche since the day you were born. The metabolic costs of task-based neural activity are low—i.e. "engaging" in a cognitive task does not drastically increase the brain's metabolic rate[2] [227,269]—*but this is because the brain is always engaged*: it must constantly generate predictions and regulate the internal milieu, even when the organism is lying still in the scanner [226]. The costs of model-based strategies, then, do not stem from activating brain regions, or searching through a decision-tree (as a computer would do); rather, they stem from a steady metabolic investment in brain structure, and informational uptake, distributed across a lifetime. However, although the costs of task-based activation are relatively small, the overall cost of the brain remains a critical concern, especially given that it consumes approximately 20% of an adult human's metabolic budget at rest [50]. Any adaptation in neural design that can minimize these ongoing costs will be advantageous [64]. In the next section, we explore a principle of neural design that controls the metabolic costs of signaling: predictive processing.

2.3. Metabolic costs of neuronal signaling are minimized by encoding prediction error

An organism implements a predictive (i.e. allostatic) model to regulate its body in its interactions with an environment [278]. Beyond minimizing errors, a predictive model can also make neural activity metabolically efficient. This efficiency is made possible by *predictive processing*, a property of signal transmission that removes redundant information. Predictive processing is a core component of information theory [266], a branch of mathematics and engineering that is central to biology, language, physics, and computer science, among other areas. For present purposes, the important point is simply that an incoming sensory signal that is perfectly predicted is redundant—it carries no information, meaning there is nothing to be encoded. For example, if a light is on then a predictive system only needs to take up information when the light is turned off (i.e. the system only encodes changes). In this way, the cost of neuronal signaling can be kept efficient by transmitting only unpredicted signals, i.e. by transmitting *prediction error*.

Neuronal signaling costs account for the majority of the brain's metabolic budget. Signaling costs account for $\sim$75% of energy expenditures in grey matter [8,260], and $\sim$40% in white matter [142]. Almost all of these costs stem from the Na^+/K^+ pump, which restores the neuronal ion gradient, extruding 3 Na^+ and importing 2 K^+ ions for each ATP consumed [8]. In grey matter—which consumes approximately three times more energy than white matter at rest [142,270]—major contributions to the signaling budget include the maintenance of the resting gradient ($\sim$11% of the signaling budget), restoration of the gradient after action potentials ($\sim$22%) and restoration after postsynaptic activations of ion channels by glutamate ($\sim$64%; [260]). Compared to these constant costs, tissue construction is a relatively minor expense [209]. If natural selection pressures organisms to economize their use of metabolic resources [64], and if adult humans devote $\sim$13% of their energy budget at rest[3] to neuronal signaling [8,50], then organisms must make signaling costs efficient to survive [38,210,259]. Predictive processing solves this dilemma, minimizing signaling costs by transmitting only signals that the internal model did not predict.

In recent years, a coherent family of mathematically formalized accounts of neural communication have emerged, with predictive processing at their core [22,23,26,41,47,48,72,100,102,156,174,230,259,262,265]. Among these accounts, the algorithmic and implementational specifics differ and are actively debated (see [275]); however, the core

---

[2] These small increases in metabolic rate during neural "activation" have been taken as evidence against resource-based accounts of cognitive effort [184,215], including a well-known account that hypothesized cognitive effort depletes circulating blood-glucose [106,105]. Evidence for this circulating blood-glucose account has also failed to replicate (e.g. [185]). But criticism of this prior work is not applicable to the present hypotheses. Allostatic accounts, like ours, maintain that vital, shared resources (like circulating glucose), are essential to survival and should not be disrupted by non-essential cognitive activity (e.g. engaging in an N-back task; [300]). Indeed, destabilizing the internal milieu is exactly what an allostatically efficient system must avoid [278,281]. The specifics of how metabolic costs are realized is an open area of research, and recent work has attempted to bridge motivation-based (e.g. [184]) and resource-based accounts, suggesting that metabolic costs might correspond to local, rather than global, metabolic changes ([312]; see also, [300]). Proposed micro-scale changes include the depletion of glycogen reserves, stored in astrocytes [44], and the accumulation of amyloid peptides [157], waste products of synaptic activity. For present purposes, our account rests only on the premises that neuronal activity is metabolically costly and that the brain is well-adapted to manage limited resources (i.e. cognitive computation is "resource-rational"; [136]; see also, [292]).

[3] The brain's metabolic consumption at rest is $\sim$20% of the body budget [50]. Of that 20%, $\sim$75% is consumed by grey matter [8], and $\sim$75% of grey matter consumption is accounted for by signaling costs (.2 *.75 *.75 = 11.25% of the body budget). White matter accounts for $\sim$25% of the brain's glucose consumption [142], of which $\sim$40% is accounted for by signaling costs (.2 *.25 *.4 = 2% of the body budget).

idea—that the brain is fundamentally predictive—is old, and is consistent with the work of Islamic philosopher Ibn al-Haytham (in his 11th century Book of Optics), Kant [165], and Helmholtz [293] (for a brief discussion, see [265]). Predictive processing approaches are also well-established in the motor learning literature [265,264,267,304], where copies of motor commands are also sent to sensory cortices (called efferent copies; [273,295]). Efferent copies modify neural activity in sensory cortices (e.g. [81,271,272,307]), allowing them to anticipate the sensory consequences of motor commands (e.g. visual, visceral, somatosensory) *before* sensory information travels from the periphery to the brain [97]. For example, people cannot easily tickle themselves [52], but when self-tickling is delayed or reoriented by a robotic hand the sensation becomes stronger [33]. That is, a predictable sensation (self-tickling) is uninformative and ignored, but when the relationship between a motor command and sensory feedback is altered (by a delay or reorientation), the efferent copy no longer predicts the sensory consequences—the sensory consequences become informative, and the sensation is experienced.

Predictive processing approaches of neural organization go further, adding that the brain is loosely organized in a predictive hierarchy [17,86,200], with primary sensory neurons at the bottom and compressed, multimodal summaries at the top. This process of prediction, comparison, and transmission of prediction error is thought to occur at all levels of the hierarchy. In general, at a given level of the hierarchy, when prediction signals mismatch with incoming information (passed from a lower level), the neurons at that level have the opportunity to change their pattern of firing to capture the unexpected input. This unexpected input is prediction error. Prediction error need not be consciously attended to be processed—its propagation is a fundamental currency of neural communication. For example, in primary sensory cortices, prediction signals are compared with incoming sensory signals (e.g. frequencies of light, pressure on the skin, etc.), whereas in association cortices prediction signals are compressed multimodal summaries of sensory and motor information, and are compared with slightly less compressed summaries of this sensory and motor information [22,41,99]. Social predictions always involve these compressed, multimodal summaries [13,16,177,213,233,282].

Predictive processing approaches have the potential to radically reorganize mainstream views of cognitive science [47] and psychological science more generally [158]. For present purposes, however, we draw two less radical conclusions: first, neuronal signaling has a metabolic cost; and second, by predicting signals (at all levels of the cortical hierarchy), and encoding only prediction error, the metabolic costs of neuronal signaling can be minimized [259].

2.3.1. Constructing and coasting

For predictive processing to be efficient, the brain must make accurate predictions in the first place. To make these predictions, the brain must encode information (i.e. encode prediction error), building on its existing model to create one that is more powerful and more generalizable. That is, to maintain metabolic efficiency in the long-run, organisms must learn. They learn by exposing themselves to novelty (e.g. [39,53]), paying a short-term metabolic cost to encode information and contribute to a model that can make accurate predictions in the future. *The goal of a brain, then, cannot be to always minimize prediction error, or to always minimize metabolic expenditures* (see, the "dark room" criticism of free energy predictive processing accounts, where the brain's primary goal is to minimize prediction error; also see, [104,262]). Instead, to survive and even thrive, organisms must invest resources wisely, managing the trade-off between the metabolic efficiency granted by their internal model's accurate predictions, and the metabolic costs of model construction, which necessarily involves taking up information as prediction error.[4]

We hypothesize that behavior involves a balancing act between these concerns. At times organisms will seek novelty (i.e. seek prediction error), *constructing* a more generalizable model of the environment. and at other times organisms will seek—or create—predictability, *coasting* on the metabolically efficient predictions of their existing model. The interplay between *constructing* and *coasting*[5] will be critical to an understanding of how humans control their social environment (for a similar approach, see [103]). Our primary concern in this paper is with a *sense of should*, which is a strategy for *coasting*—we will suggest that conforming to others' expectations creates a predictable social

---

[4] An account of how organisms make decisions that balance constructing and coasting is beyond the scope of this paper. Such an account would require a general theory of value and decision-making, perhaps where value is metabolically defined—e.g. highly valued decisions are anticipated to be metabolically advantageous over some flexible time-horizon.

[5] Constructing and coasting are akin to exploring and exploiting (e.g. [53]). As concepts, exploring and exploiting emphasize an organism's behavior. By contrast, we use constructing and coasting to emphasize an organism's internal model, and the strategic benefits of changing or maintaining that model.

environment, minimizing the metabolic costs of prediction error (but see Section 3.4, for an example of *construction* in the context of mental inference).

2.4. Summary

A brain implements a predictive model to regulate an organism's body in its environment [54,236,278,281]. A brain is also a significant metabolic investment [50], and as organisms must be metabolically efficient to survive, the brain's energetic costs must be regulated (especially the high costs of neuronal signaling; [8]). Predictive processing satisfies this need for neuronal efficiency by limiting energy expenditures, transmitting only unpredicted signals from one level of the neural hierarchy to the next [47,100,266]. It follows then, that prediction error carries a metabolic cost [259], and unpredictable environments are metabolically costly.

From this empirical foundation, we can develop our account of a *sense of should*. This account hinges on one additional point: *you contribute to the social environment of others, and they comprise the social environment for you*. Encoding information (i.e. encoding prediction error) about these other people is a metabolic demand, but this metabolic demand can be controlled. We propose that humans learn to control the behavior of others (and by extension, the metabolic demands others impose) by conforming to their expectations. This control is not coercive—that is, others are not forced to perform particular behaviors—rather, this form of control can make others' behavior more predictable. Other people can be made more predictable when you are predictable to them.

3. A metabolic and predictive framework for modeling a *sense of should*

In this section, we outline our central hypothesis: that a *sense of should* regulates the metabolic pressures of group living, i.e. that people are motivated to conform to others' expectations, and by conforming, they maintain a more predictable—and by extension, a more metabolically efficient—social environment. If your behavior conforms to other people's predictions (i.e. if your behavior minimizes prediction error for them) then they will have less reason to change their behavior, making them more predictable for you.

We suggest that a *sense of should* serves a metabolic function, and that it should develop in nearly all humans—but we are not assuming that it is innate. On our account, there is no need to assume that a *sense of should* is a specialized, or domain-specific adaptation (cf. [56]). Rather, we suggest that a *sense of should* is an emergent product, both of domain-general capabilities (e.g. [150]) that are exceptionally well-developed in humans (e.g. associative learning, memory), and of social context (specifically, a social context where others' behaviors are contingent on your own). Further, the metabolic benefits of a *sense of should* almost certainly coexist (or conflict) with other adaptive strategies, including self-interested hedonically motivated behavior, exploration, reputation-seeking, or reciprocal altruism in repeated interactions (e.g. [11,288]). In this section, we develop a formal model of a *sense of should* (using mathematical formalism to make all assumptions explicit) and in section 4, we elaborate on the implications of this model in dynamic social contexts. Our story, then, begins with metabolic frugality, but ends with the complex interplay of motivations that characterize human social life.

3.1. The metabolic benefits of conformity

To formalize the individual adaptive benefits of conforming to others' expectations, we use a working example: a person named Amelia. We assume that Amelia's brain, like the brain of any organism, consumes metabolic resources to maintain her internal milieu and to move her body around the world. Amelia's brain processes unexpected sensory information as prediction error, which is neurally communicated at a metabolic cost (section 2.3). Formally:

$$M_{total} = M_{pe} + M_{other} \tag{1}$$

where

M_{total} represents Amelia's total metabolic expenditures across some arbitrary time period,

M_{pe} represents the metabolic costs of encoding prediction error across that time period, and

M_{other} represents other metabolic costs not related to neuronal signaling.

In predictive processing models, a precision term, weighting prediction errors according to their certainty, is often included (e.g. [83]), but for the sake of simplicity we omit these terms while developing our model (but see section 4.1.1).

Prediction error comes from sensory changes in the body (interoceptive sources) and sensory changes in the surrounding world (exteroceptive sources). Interoceptive prediction error refers to unexpected information about the condition of the body (signaling, for example, heart-rate, respiration, metabolic and immunological functioning; [26, 58,261]). Exteroceptive prediction error refers to unexpected information in the environment (signaled by sights, sounds, etc.). Exteroceptive prediction error, experienced by Amelia, could come from many sources, each of which could be defined as an entity[6] (e.g. animals, machines, inanimate objects, the weather). For the purposes of our model, the critical distinction among entities is *whether a given entity does, or does not, predict Amelia's behavior*. If an entity predicts Amelia's behavior, then the prediction error she receives from that entity is called *reciprocal prediction error* (our examples assume that these entities are human, but see footnote 17 for an extension to non-biological entities). If an entity does not predict Amelia's behavior (e.g. as in weather, falling rocks, walls, ceilings), then the prediction error she receives from that entity is called *non-reciprocal prediction error*. Formally:

$$M_{pe} \propto pe^{Int} + \sum_{i=1}^{n} pe_i^{Ext:R} + \sum_{i=1}^{m} pe_i^{Ext:\sim R} \tag{2}$$

where

M_{pe} represents the metabolic cost (to Amelia) of encoding prediction error across a time period,

pe^{Int} represents Amelia's interoceptive prediction error,

$\sum_{i=1}^{n} pe_i^{Ext:R}$ represents Amelia's reciprocal prediction error, from n entities in the environment,

$\sum_{i=1}^{m} pe_i^{Ext:\sim R}$ represents Amelia's non-reciprocal prediction error, from m entities in the environment, and

$\propto$ denotes a proportional relationship, as the exact relation between prediction error and metabolic cost is unknown. For Amelia, the adaptive advantage of conformity stems from regulating reciprocal prediction error.

Conforming to others' expectations benefits Amelia by reducing the likelihood that others will change their behavior in unanticipated ways—i.e. all else being equal, conforming keeps others more predictable. This conclusion can be derived by examining reciprocal prediction error. The reciprocal prediction error experienced by Amelia is generated by multiple entities in her environment, but for now we narrow the focus to one person, named Bob. Using her internal model, Amelia predicts Bob's behavior, and her prediction error from Bob ($pe_{A:B}^{Ext:R}$) equals the magnitude of the difference between Bob's behavior (b_B) and her prediction about his behavior ($pb_{A:B}$).

$$pe_{A:B}^{Ext:R} = |b_B - pb_{A:B}|$$

Likewise, Bob predicts Amelia's behavior, and his prediction error ($pe_{B:A}^{Ext:R}$) equals the magnitude of the difference between Amelia's behavior (b_A) and his prediction about her behavior ($pb_{B:A}$).

$$pe_{B:A}^{Ext:R} = |b_A - pb_{B:A}|$$

When Amelia's behavior deviates from Bob's predictions (i.e. when $|b_A - pb_{B:A}| > 0$) Bob receives information in the form of prediction error [266]. This information may cause some change to Bob's internal, predictive model (X_B), proportional to the amount of information provided. Critically, if Bob encodes the prediction error (i.e. Bob learns), then these changes in Bob's internal model may cause a proportionate change in his behavior (b_B). That is, on average, when Bob's predictions are violated, Bob may change his internal model by some amount, and his behavior may change with it.

$$\Delta b_B \propto \Delta X_B \propto |b_A - pb_{B:A}|$$

If Bob's behavior changes, and if Amelia is unable to anticipate *exactly* how it will change (in the next moment, and in some number of moments following it), then prediction error will increase for Amelia.[7]

$$pe_{A:B}^{Ext:R} = |b_B - pb_{A:B}| \propto \Delta b_B$$

---

[6] How entities are identified is a deeper problem for cognitive psychology and neuroscience (and the social sciences more generally; [80]). For modeling purposes, we assume that it can be done, with the caveat that this process of segmentation itself may be affected by many factors.

[7] Under some circumstances Amelia may accurately predict how Bob's behavior will change after she violates his predictions (e.g. if she knows something about his internal model, X_B). But typically, it will be easier for Amelia to predict Bob's behavior by conforming to his expectations, as in this case she could simply predict that Bob will continue doing what he was doing previously.
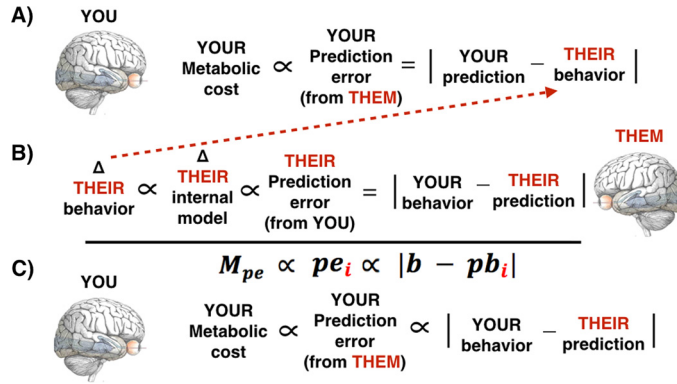
Fig. 2. Illustration of the derivation of Equation (4), modeling the control of prediction error (and its metabolic costs) by conforming to others' predictions. **A)** Your prediction error equals the difference between the predicted and actual behavior of another person, and is assumed to carry a metabolic cost (section 2.3). Others' predictions can be understood as a vector of sensory signals, and your behavior is a matched length vector. **B)** Prediction error for others equals the difference between your behavior and their prediction. As prediction error is informative [266], prediction error produces some proportional change in others' internal models, which in turn produces some proportional change in their behavior. **C)** Considered together, the relationships imply that your prediction error (and its metabolic costs) are more likely to increase when your behavior violates others' predictions.

It follows, then, that the prediction error Amelia experiences from Bob ($pe_{A:B}^{Ext:R}$), is related to the difference between her behavior and Bob's predictions about her behavior ($|b_A - pb_{B:A}|$). This relationship is mediated by changes in Bob's internal model and his behavior (Fig. 2). Formally:

$$pe_{A:B}^{Ext:R} = |b_B - pb_{A:B}| \propto \Delta b_B \propto \Delta X_B \propto |b_A - pb_{B:A}| \tag{3}$$

which reduces to:

$$pe_{A:B}^{Ext:R} \propto |b_A - pb_{B:A}|$$

Thus, Amelia can manage the metabolic costs imposed by Bob by conforming to his expectations. In the more general case, Bob is one arbitrary entity (i):

$$M_{pe} \propto pe_i^{Ext:R} \propto |b - pb_i| \tag{4}$$

where

　　M_{pe} represents the metabolic cost (to Amelia) of encoding prediction error across a time period,
　　$pe_i^{Ext:R}$ represents Amelia's reciprocal prediction error, from one entity (i) in the environment,
　　b represents Amelia's behavior, and
　　pb_i represents the prediction of one entity (i) about Amelia's behavior.

Thus, the prediction error experienced by Amelia, from one entity (e.g. Bob), and the metabolic costs of that prediction error, are proportional to the discrepancy between her behavior and Bob's predictions about her behavior. In this formulation (which is intentionally presented as a simplified sketch of the core components; a full version would include, at a minimum, precision weighting terms; see section 4.1.1), Amelia's behavior (b) could be treated as a binomial vector, representing all possible features of her behavior in a given instance (i.e. each entry in the vector denotes one specific feature of her behavior, marking it as present or absent). A prediction about her behavior (pb_i) is a matched length binomial vector, meaning Bob predicts the features of Amelia's behavior. When these two vectors are identical, both the proportional change in Bob's behavior, and the proportional increase in Amelia's prediction error, are minimized.[8] All else being equal, Amelia can regulate the metabolic costs of prediction error by conforming to others' expectations.

---

[8] Attention might be integrated into this model by altering the length or specificity of vector pb_i. For example, if Amelia's behavior is scrutinized by an entity, then the vector pb_i (and its matched vector, b) might contain more entries, increasing the potential maximum of $|b - pb_i|$.

A brain implements a predictive model to regulate an organism's body in its environment ([54,236,278,281]; see Section 2.2). With respect to this model, organisms must balance (at least) two pressures: they must balance *constructing* (e.g. disrupting their environment at some metabolic cost, but gaining information that can be added to their internal model and inform future predictions) and *coasting* (e.g. using their existing internal model to make accurate, metabolically efficient predictions; section 2.3.1). Broadly, a *sense of should* is a strategy to facilitate *coasting*—it maintains a predictable social environment, allowing Amelia's internal model to continue issuing accurate predictions. To issue accurate predictions, Amelia had to invest in *constructing* a sufficiently accurate internal model in the first place. In this way, conforming to others' expectations secures Amelia's initial investment, as every time that she violates others' expectations, she increases the likelihood that their behavior (and their internal models) will change in ways that she cannot predict. If other people changed in ways Amelia could not predict, then she would need to reinvest in model construction all over again. As long as she conforms to others' predictions (and as long as others' internal models don't suddenly, and unbeknownst to Amelia, change dramatically), her predictions of others will remain relatively accurate and efficient.

This process—conforming to others' expectations in order to *coast* on a predictable environment—also implies a positive feedback loop: when Amelia conforms, she makes her environment more predictable, which will make mental inference (i.e. estimating pb_i; see section 3.4) easier, which then makes conforming easier still. But this positive feedback loop cannot run forever: eventually, overconforming may come at a cost to Amelia's survival or reproduction. That is, at an extreme, Amelia might become a doormat.[9] *Coasting* on the metabolic benefits of a *sense of should*, then, must be balanced with satisfying other adaptive needs for survival and reproduction. Amelia cannot only conform to social pressure, she must balance the benefits of a predictable social environment against her other needs (e.g. food, sex, safety).[10]

But the predictable social environment generated by Amelia's conformity may also help her implement other strategies, ranging from deception to reciprocal altruism. If Amelia's social environment is predictable, and if her behavior ensures that it will remain predictable, then she can engage in long-term social planning. This social planning could be cooperative or competitive: Amelia can leverage the social predictability (that she helped create) to extend alliances with other people, or she can use social predictability identify occasions where it is to her advantage to deceive or betray them. Thus, although a *sense of should* is experienced as a motivation distinct from self-interested reputation-seeking, or utility maximization ([7, Chapter 12]; [27,75,133,221,251,268,286]), its interplay with other motives within a social environment can make new strategies possible.

This section has shown how it could be individually adaptive to conform to others' expectations, and how these advantages follow from the biological foundations already established in section 2. In the following sections, we develop the framework surrounding this model further, showing how a *sense of should* is experienced as a psychologically distinct motivation (section 3.2), how it might develop, (section 3.3), how it facilitates mental inference (section 3.4), and how social influence via a *sense of should* differs from social influence as it is more typically considered (section 3.5).

3.2. The psychological experience of a sense of should

We have established that a *sense of should* can regulate predictability in a social environment, and that this strategy is distinct from the pursuit of reputation or social reward. But Adam Smith [268], and others (e.g. [7,27,75,133,221, 251,286]) go further, suggesting that a *sense of should* is psychologically distinct from a more general motivation to seek rewards, such as reputation. Following them, we suggest that unpredictability can be aversive *in and of itself* [154,155]. When Amelia violates others' expectations, she disrupts her social environment, producing metabolic and affective consequences. When this relationship between cause and consequence is learned, Amelia's brain should

---

[9] However, individual differences in the reward value of social predictability may exist, providing a natural point for our framework to interface with theories of individual differences (e.g. personality).

[10] However, the costs of violating close others' expectations may still loom large. Each human is embedded in a web of social relationships (as emphasized in relational sociology approaches; [80]), and these social relationships can be considered in terms of the expectations of each member (e.g. you have expectations about your best friend, your mentor, your boss, your romantic partner, and they all have expectations about you). As so much of your life depends on maintaining these relationships (and satisfying the expectations that comprise them), social pressure will most likely remain a major source of motivation, even when it conflicts with other goals. Indeed, violating others' expectations in major ways (e.g. betraying a family member) could affect many of these relationships at once, rearranging the entire network of expectations and disrupting your social niche (at a significant metabolic cost).

motivate her to regulate these violations of others' expectations prospectively (i.e. allostatically; section 2.2), allowing a *sense of should* to emerge as an anticipatory aversion to violating others' expectations. To provide a full account of this process, we briefly review the modern scientific understanding of affect.

Affect refers to the psychological experience of valence (i.e. pleasantness vs. unpleasantness) and arousal (i.e. alertness and bodily activation vs. sleepiness and stillness). Valence and arousal are core features of consciousness [22,63,77,79,159,254,256,306], and, when intense, become the basis of emotional experience ([20,24,240,241], [306, Chapter 7]). Affect is a low-dimensional transformation of interoceptive signals, which communicate the autonomic, immunological, and metabolic status of the body [22,24,26,41,58,261,263]. Valence and arousal are sometimes considered to be independent dimensions of affect, but, in reality, they exhibit complex interdependencies [24,95,124, 182]. For present purposes, it is enough to say that arousal is not necessarily valenced, yet, context will guide its interpretation as pleasant or unpleasant [22,23].

Recent work has demonstrated that prediction error is associated with the physiological correlates of arousal. For example, prediction error is associated with electrodermal, pupillary, neurochemical, and cardiovascular responses that reflect patterns of ANS (autonomic nervous system) arousal [36,59,61,68,140,194,224,276,309]. Unpredictable environments, then, including ones created by Amelia's non-conformity, generate arousal, and this arousal will be interpreted in the context of ongoing exteroceptive and interoceptive information. Given this, we can assert that unpredictable social environments are arousing; what must also be established, is how they become aversive.

Having one's expectations violated is sometimes pleasant, and sometimes aversive. For example, comedy often stems from incongruity and transgressing norms [49,212]. Likewise, intentionally provoking a speaker with a pointed question may be disruptive, but their answer could be informative (and therefore useful for *constructing* the brain's internal model, facilitating future predictions; section 2.3.1). Disruptions can be adaptive. However, disruptions will always involve processing prediction error, and therefore, they *will always be metabolically costly* (section 2.3). Given this, in the absence of some other benefit to be gained, metabolic efficiency is best served by avoiding such disruptions, i.e. *coasting* on the brain's existing model. Further, violating others' expectations is risky. For example, if Amelia tells a dirty joke, how others interpret their arousal will decide whether the joke is interpreted as hilarious or offensive. Thus, although it can occasionally be pleasant to violate other's expectations, there is reason to expect that transgressing norms will often be stressful and unpleasant.

When Amelia violates others' expectations, then, she invites an aversive outcome (i.e. "punishment"). But for a *sense of should*, the "punishment" does not come from other people, or at least, not explicitly from them—no second or third party intentionally administered it for the purpose of punishing Amelia, nor did anyone pay a cost or risk anything to censure her. Instead, for Amelia to receive the punishment, it is only necessary that others react naturally to their expectations being violated, changing their internal model, and changing their behavior with it (Equation (4); Fig. 2). When others' behaviors change, prediction error increases for Amelia, and the metabolic efficiency of her internal model temporarily suffers. She will experience arousal, which if intense or pervasive enough will often be experienced as aversive. The punishment, then, arrives as both a metabolic cost, and as the experience of negative affect. The affective experience was not something imposed on Amelia by others; rather, it stems from the way she makes meaning of her own interoceptive sensations [22,23]. Indeed, classic accounts of helping behavior (a special case of conforming to others' expectations) suggest that helping arises from the combination of evoked arousal by a suffering victim, and the helper's ability to reduce that arousal by helping [75,221]. The categorization of interoceptive sensation was even present in classic accounts of moral development:

> "Two adolescents, thinking of stealing, may have the same feeling of anxiety in the pit of their stomachs. One . . . interprets the feeling as 'being chicken' and ignores it. The other . . . interprets the feeling as 'the warning of my conscience' and decides accordingly." [176, pp. 189–190]

This, then, may be what separates a "self-interested" motivation to pursue rewards and avoid punishments [268], from a *sense of should* (and possibly from moral obligation, which is beyond our present scope; Fig. 1). Rewards and punishments (e.g. pleasure, pain) are externally administered, whereas a *sense of should* necessarily involves self-caused disruptions of the social environment, and a subsequent interpretation of interoceptive sensation. It is a punishment that Amelia's brain literally inflicts on itself.

It must be noted that Amelia experiences aversive outcomes through her interpretation of affect, but a *sense of should* actually motivates her to avoid such consequences prospectively. That is, we propose that a *sense of should*

is experienced as an *anticipatory* aversion to violating others' expectations.[11] The brain is an allostatic regulator—it anticipates the needs of the body and attempts to meet those needs before they arise, thereby avoiding errors (section 2.2). If Amelia's brain has learned that violating others' expectations decreases her metabolic efficiency (and consciously, Amelia experiences violating others' expectations as aversive), then, Amelia will prospectively avoid such situations and the behaviors that trigger them. In the absence of some competing goal, the most metabolically efficient option will often be to behave as others expect. A *sense of should*, then, is not an exceptional motivation—in social settings, a *sense of should is a default*. Given the metabolic importance of a predictable social niche, and given that any individual can disrupt that niche by violating others' expectations, we hypothesize that, all else being equal, adults continuously adjust their behavior to fit others' expectations, only rarely making a hard break from observing social norms to exclusively pursue their own interests. Indeed, if this weren't true, group living might be impossible.

3.3. The development of a sense of should

A *sense of should* is a motivation to prospectively avoid behaviors that deviate from others' expectations in the service of metabolic efficiency. However, a *sense of should* does not involve conditioning avoidance of any specific behavior; rather, it involves learning a *relationship*. The relationship is between your behavior and others' expectations, $|b - pb_i|$. When the discrepancy between your behavior and others' expectations is large, the social environment becomes less predictable, and metabolic and affective consequences follow. Put another way, developing a *sense of should* involves learning what behaviors are *appropriate* (i.e. expected by others) in a given context.

To learn this relationship, Amelia must accomplish at least two developmental tasks. First, she must be able to accurately predict the behaviors of social agents (as otherwise she cannot experience prediction error when her predictions are violated). Her ability to make sophisticated predictions, especially ones that extend beyond the present, will develop gradually during infancy and early childhood. Newborns live in an environment structured by their caregivers, and the predictions necessary for a newborn's survival are largely limited to those dyads—e.g. newborns learn that crying summons a blurry shape (i.e. a parent) that relieves interoceptive discomfort by feeding, burping or hugging them [9]. As Amelia's newborn brain develops a more sophisticated internal model, and as it begins to initiate interactions with adults and other children, it constructs increasingly sophisticated predictions about their behaviors. The more frequent and sophisticated these predictions are, the more potential there is for them to be violated, and for their metabolic/affective consequences to be experienced.

How these metabolic and affective consequences motivate behavior may also change across the lifespan. For example, in childhood the brain accounts for a larger portion of the whole body metabolic budget [131,169], meaning that, compared to adults, children may be more likely to tolerate fluctuations in the metabolic costs imposed by their environment, as these fluctuations comprise a smaller portion of the brain's total metabolic budget. Likewise, older adults are more likely to self-select out of high-arousal situations [245,246], and are more likely to experience high arousal stimuli as aversive (regardless of whether the stimuli were experienced as positive or negative by younger adults; [166]) suggesting that older adults may be less likely to tolerate these metabolic costs (or the corresponding affective experiences). Further, young children (e.g. 4-year olds) are more likely to entertain a range of predictions (i.e. an explore strategy, which would oppose a *sense of should*) and adults are more likely to limit predictions to the outcomes that are most likely (i.e. an exploit strategy, which a *sense of should* facilitates; [126,127,190,258]), potentially as a consequence of the late development of prefrontal cortex and associated processes supporting cognitive control [284]. In the context of social pressure, our framework suggests that these developmental changes—e.g. in metabolic efficiency, sensitivity to arousal, and cognitive control—may underlie changes in sensitivity to a *sense of should*, and that, as a social consequence of these changes, children may show less aversion to unpredictable social settings, whereas older adults may strive to maintain this social stability.[12]

---

[11] As the aversion is anticipatory, it is also possible for Amelia to be wrong about others' expectations, and about the consequences of her nonconformity. This raises some exciting avenues for research in social anxiety: some people may pathologically overestimate how severely their behavior will disrupt their social environment, or overweight aversive interoceptive experience [170].

[12] Sensitivity to social pressure and to social stability during adolescence is a complex topic, and well beyond our present scope; however, it is worth noting that adolescents adopt more exploratory learning strategies in social settings [127], but at the same time, are highly influenced by the judgments of others [31]—influence that is mediated by BOLD activity in regions supporting allostasis and interoception [174]. Further, developmental changes during adolescence may cause experiences associated with stress to be "longer lasting and qualitatively different from

The second developmental task is for Amelia to develop the ability to make precise inferences about others' expectations of her. A *sense of should* involves learning a relationship between her behavior and others' expectations ($|\boldsymbol{b} - \boldsymbol{pb_i}|$), and Amelia must infer others' expectations ($\boldsymbol{pb_i}$) precisely enough to identify when her behavior conforms, and when it is discrepant. In the next section (section 3.4), we outline how this capability might develop. We hypothesize that Amelia is born with a minimal "toolkit" of domain general processes (e.g. memory, associative learning; for a similar view, see [150]), and from this foundation, develops a fine-tuned ability to make inferences about the predictions made by others' internal models (i.e. an ability to engage in mental inference). Given this, children may experience arousal in unpredictable social settings, but it may only be later in development that they understand that the relationship between their own behavior and others' expectations regulates this arousal, and only when they learn this contingency will they feel obligated to conform to others' expectations (for a similar account of empathic development, see [153]; for review, see [75]).

3.4. Mental inference and a sense of should

Inferences about others' expectations are at the core of our approach (Equation (4); Fig. 2). To select a behavior that matches others' expectations, and that controls prediction error in the social environment, Amelia must first infer what behavior others expect of her. We use the term *mental inference* to stand in for all of these inferences about others' expectations, with the caveat that others' expectations may be formulated as high-level, abstract predictions about mental states, as low-level, concrete predictions about behaviors, or as predictions at any level of abstraction in between [178,291]. There are many competing accounts of mental inference, and most likely a number of underlying proficiencies and/or cognitive processes that combine to facilitate it [4,116,247,296], but the core problem that accounts of mental inference aim to solve is this: *How do people make inferences about others' minds (i.e. predictions generated by others' internal models),*[13] *given that others' internal models cannot be directly observed.* Three prominent theoretical perspectives—simulation theories, modular theories, and 'theory' theory—all provide different answers. Simulation theories suggest that Amelia performs mental inference by using her own mind (i.e. her internal model) as a simulator (e.g. [122,123,130]). For example, she may feed "pretend beliefs" and "pretend desires" into her own "decision-making mechanism", treating the "output" as the inferred mental state [123, p. 452]. Modular theories propose that mental inference is made possible by innately specified cognitive mechanisms (e.g. [14,18,186,187,248]), claiming, for example, that "the concepts of belief, desire, and pretense [are] part of our genetic endowment", and that mental inference is made possible by "a module that spontaneously and postperceptually processes behaviors that are attended, and computes the mental states that contributed to them" [248, p. 697]. Finally, 'theory' theory proposes that mental inference is a subcategory of the more general process of inference [125,128,129]. That is, in mental inference, as in learning more generally, children construct theories: they "infer causal structure from statistical information, *through their own actions on the world* and through observations of the actions of others" [emphasis added] [129, p. 1085]. Adjudicating between these accounts is beyond the scope of this paper, but our approach can make clear how domain general trial-and-error learning (as in 'theory' theory), combined with the use of prior information (as in simulation theory) might allow one brain to make inferences about the unobservable predictions of another. We suggest that mental inference is necessary to experience a *sense of should*, and that conversely, the interpersonal

---

stress exposure at other periods of life, possibly due to the interaction between the developing hypothalamic-pituitary-adrenal (HPA) axis and glucocorticoids" ([34, pp. 189–190]; for review, see [195]). How culture and development combine to shape reactions to social pressure during adolescence will be a difficult problem to solve, but our framework can be used to structure hypotheses aimed at addressing this question.

[13] Most mainstream accounts of mental inference are interested in explicit inferences about others' propositional beliefs, desires, or intentions—i.e. mental representations (for review, see [4,310]). Having knowledge of these mental representations (and reasoning about them) is typically referred to as having a "theory of mind" [223]. However, recent work has shown poor correlations in theory of mind measures across development [296], and cross-cultural work has demonstrated that explicit inferences about mental states are most frequent in Western societies [78,115,198]. In this paper, we avoid the term "theory of mind" and its implication that mental states are propositional representations, instead using "mental inference" to refer more generally to inferences about others' predictions about Amelia's behavior at all levels of abstraction (i.e. from abstract mental states to concrete features of action). The process of mental inference may sometimes produce articulatable, explicit expectations about Amelia's beliefs, preferences and emotional experience—made articulatable via culturally inherited concepts [23,150]—however, for present purposes, our focus is on how the process of mental inference relates to general cognitive processes (e.g. memory, associative learning) and their emergent properties as these general cognitive processes are shaped across development within a social environment.

dynamics that make a *sense of should* possible (Equation (4); Fig. 2) can be used to facilitate more precise mental inferences.

As discussed in section 3.1, if Amelia's behavior violates someone's expectations (e.g. Bob), then Bob's internal model and behavior will change proportional to the violation. This change in Bob's behavior creates prediction error for Amelia (Equation (4); Fig. 2). There is a relationship, then, between Bob's predictions about Amelia (which are generated by his internal model), and the prediction error that Amelia receives from him. Amelia cannot infer *exactly* what Bob predicts, but she can identify when she has violated Bob's predictions: when Amelia has violated Bob's predictions, his behavior is more likely to change, increasing prediction error for her. This link—between others' predictions and the prediction error Amelia receives when violating them—may provide a route through which Amelia can cumulatively construct a model of others' minds. Further, by using this route in combination with her prior knowledge about Bob, or people more generally, Amelia can inform her guesses about what Bob's predictions might be, reducing the need for metabolically expensive trial-and-error learning. We demonstrate this below, extending our model from section 3.1.

Prediction error experienced by Amelia, from one person (i), is proportional to the discrepancy between her behavior and his prediction.

$$pe_i^{Ext:R} \propto |b - pb_i|$$

But, Amelia has no direct access to his prediction. Instead, she must infer it. The equation can be rewritten to only include information accessible to Amelia: her prediction error, her behavior, *her mental inference* about what someone expects, and the error in that inference. Initially, the error in Amelia's inference will be unknown to her, but we will suggest that she can estimate it by applying prior knowledge and engaging in a dynamic process of trial and error—forming a mental inference, enacting a behavior, then estimating her error.

$$pe_i^{Ext:R} \propto |b - (pb_i^M + e)| \tag{5}$$

where

pb_i^M is a vector representing Amelia's estimate (i.e. her mental inference) of entity i's prediction about her behavior, and

e is the error in Amelia's estimate, such that $pb_i^M + e = pb_i$.

For example, Amelia cannot directly confirm that her father expects her to call him on his birthday, as his expectations are not externally observable. However, she can infer, based on her prior knowledge, that he probably expects a call. In this case, Amelia can use her inference ($pb_i^M + e$) to stand in for the actual predictions her father has about her behavior (pb_i). There is always the possibility that she is wrong (i.e. that e is large). For example, she may have accidentally offended him the day before, and he may prefer that she not call this year.

Amelia had to use prior knowledge to generate her inference about her father's birthday expectations. The prior knowledge informing her prediction could come from many sources, but most obviously, it could come from her prior experience with her father. For example, if she knows he expected a call last year (or even that he is sensitive and cares about this sort of gesture in other non-birthday contexts), then she has some reason to infer that he expects a call today. To provide a formalized sketch of this route to inference, we represent Amelia's estimate of her father's current prediction (pb_i^M) as a Bayesian posterior, conditioned on some number (n) of prior predictions she knows he has made (pb_{it}). In psychology, such an inference about a particular person is commonly called a *dispositional inference* ([144,161,168]; for review, see [117,191]).

$$P\left(pb_i^M \middle| \sum_{t=0}^{-n} pb_{it}\right)$$

At another extreme, Amelia could use prior knowledge from other people (aside from her father) to generate an inference about what her father expects. For example, she could infer that her father expects a phone call through her prior experience with *everyone* who has had a birthday. This is a complementary route to the same inference. Here, the context (i.e. it being someone's birthday) is held constant, and Amelia infers her father's expectation using her knowledge of others' predictions in the same context. Again, we represent Amelia's estimate of her father's prediction (pb_i^M) as a Bayesian posterior, this time conditioned on the average of some number (n) of prior predictions (pb_{it})

that some number of entities (m) have made. In psychology, an inference based on what people typically do (i.e. what they do on average) within a given context is commonly called a *situational inference* [117,168].

$$P\left(pb_i^M \middle| \frac{\sum_{i=1}^m \sum_{t=0}^{-n_i} pb_{it}}{\sum_{i=1}^m \mathbf{n}_i}\right)$$

As both dispositional and situational inferences use prior experience, they should be imprecise in infancy and early childhood, but gradually become more refined as children grow and accumulate experience (section 3.3).[14] In this way, as Amelia's internal model takes on more information across development, it increases the amount of prior information on which her predictions can be based, akin to the core insight of simulation theory [122,123,130].

This approach (and others, see [13]), can articulate how dispositional inferences, situational inferences, and other combinations of prior knowledge are used to estimate the predictions others might make. That is, all of these forms of inference are special cases of the more general process of applying prior knowledge. For example, Amelia's *situational inference* used knowledge about all entities (m) in a given context to generate an inference about her father's predictions. She could also have used some subset of m (e.g. other fathers, other men, or other older adults). This provides a natural means of integrating stereotypes into our approach, as such subsets might also be formed on the basis of observable features (e.g. skin color, accent; [172]) or other grouping factors that Amelia's internal model has learned to see as relevant.[15]

However, as an explanation of mental inference, an account that only used prior experience to guide inference could not be complete. Such an account would be circular: all means of estimating of others' expectations (pb_i^M) would require prior estimates of others' expectations. That is, the examples of *dispositional* and *situational* inference reviewed above have required that Amelia use what others expected *before* (pb_{it}) to estimate what they expect *now* (pb_i^M). Assuming that Amelia has no innate knowledge of what others expect, such a model cannot answer how Amelia ever formed an estimate about others' expectations in the first place. We've arrived at the same hurdle as all other accounts of mental inference: *if other minds cannot be directly observed, then how can Amelia infer their contents (i.e. their predictions)?*

As alluded to at the beginning of this section, we hypothesize that Amelia can learn about others' predictions by violating them. She cannot determine precisely what others expect of her, but she can use prior knowledge to form an estimate (even an imprecise one), enact a behavior, and then use the resultant prediction error to determine whether her estimate was accurate. Extending the analogy from 'theory' theory, where mental inference is built from a process similar to scientific inference [125,128,129]: Amelia's estimate (pb_i^M) could be considered as a *hypothesis* about an entity's expectation, her behavior (b) could be considered as an *experiment*, and the prediction error ($pe_i^{Ext:R}$) generated by the entity for Amelia (including her associated metabolic costs and affective experiences; section 3.2) could be considered as *evidence*. Through this process, Amelia can estimate the inaccuracy of her initial estimate (e). Formally, given Equation (5):

$$pe_i^{Ext:R} \propto |b - (pb_i^M + e)|$$

If b and pb_i^M are known, then:

$$pe_i^{Ext:R} \propto e$$

where

$pe_i^{Ext:R}$ is the prediction error experienced by Amelia from an entity (e.g. Bob), and

e is the error in Amelia's estimate of Bob's prediction about her behavior (i.e. the error in her mental inference).

If Amelia iteratively forms hypotheses, enacts behaviors, and updates her hypotheses according to the evidence (i.e. according to the prediction error received from changes in Bob's behavior), then she can gradually infer Bob's predictions about her through his reactions to her behavior (and, potentially, through her affective experience of the resultant prediction error, consistent with the suggestion that interaction and embodiment are crucial components of mental inference; [69,93,107–110]). If, across iterations, Bob's behavior becomes more predictable for Amelia (and

---

[14] Intriguingly, this opens inroads to connect memory research with mental inference, as greater access to prior experience implies that mental inference can be made increasingly precise.

[15] How exactly entities and contexts are grouped or deemed relevant for inference is a larger, and more fundamental question for cognitive science.

her arousal decreases), then Amelia's behaviors are more likely to be approaching convergence with Bob's predictions. If, across iterations, Bob's behavior becomes less predictable for Amelia (and her arousal increases), then Amelia's behaviors are more likely to be diverging from Bob's predictions. On each iteration, Amelia's brain is using prediction error (and possibly affect) to estimate the error in her previous hypothesis about what Bob predicted, which in turn, allows her to generate a new, more accurate hypothesis. Through this cumulative process, Amelia may construct inferences about others' unobservable predictions. We refer to this route to mental inference—where estimates are created, behaviors are enacted, and evidence is evaluated—as *interactive inference* (after Shaun Gallagher's interaction theory, where understanding others is understood, in part, as an embodied practice; [107,108]).

Mental inference, then, may involve the coordinated usage of a collection of proficiencies and cognitive processes [4,116,247,296]. Interactive inference may be one component of mental inference, but we hypothesize that it works in conjunction with (at a minimum) prior knowledge, such as dispositional and situational inferences (and their permutations, e.g. stereotypes). That is, we hypothesize that Amelia uses her prior knowledge of individuals, contexts, and combinations thereof to narrow the scope of potential hypotheses.[16] With this scope narrowed, she can make the process of interactive inference efficient: choosing an estimate from this limited hypothesis space, enacting a behavior, and fine-tuning subsequent estimates and behaviors accordingly.[17]

In section 3.1, we suggested that a *sense of should* is a strategy for *coasting* on the predictions of an internal model. It maintains the predictability of the social environment, facilitating social prediction and reducing the metabolic costs of prediction error. In this section, we have outlined how the same relationship—$pe_i^{Ext:R} \propto |b - pb_i|$, linking Amelia's behavior, others' expectations, and prediction error in her social environment—may facilitate mental inference, allowing Amelia to *construct* her internal model of others' predictions about her. By using prior knowledge (e.g. dispositional and situational inferences) to guide her initial hypotheses, she can perform controlled "experiments" via *interactive inference*, fine-tuning her internal model's estimates of the minds of others. As survival depends on both *coasting* and *constructing*, no one strategy can dominate. At times, Amelia must make a metabolic investment in exploration, violating others' expectations to *construct* a more precise model of the world; and in turn, these investments allow her to more easily exploit the environment later, minimizing metabolic costs by *coasting* on her model's accurate predictions. A *sense of should* is a strategy to keep the social environment predictable, maintaining the social conditions on which Amelia's predictions depend and securing the metabolic investments in *construction* that she has already made.

3.4.1. Interaction may facilitate precise mental inference, and precise mental inference may facilitate social cohesion

Our aim in this paper is to describe how a *sense of should* is adaptive, how it is experienced, and how it is made possible by a process of mental inference that leverages domain-general mechanisms. A more detailed account of mental inference would go beyond our present scope; but, there are two points that are important enough to be worth emphasizing.

First, our approach to modeling mental inference necessarily involves interaction [107,108,110]: Amelia cannot fine-tune her inferences about others' expectations without engaging in interactive inference. An imprecise, initial

---

[16] This process of narrowing the hypothesis space then exploring it, may be especially necessary for living entities, as they are dynamic systems [274]: their internal models undergo complex changes over time, which by extension changes the predictions that they make. Note, however, that our definition of "reciprocal prediction error" (section 3.1) technically encompasses prediction error from predicting—but non-biological—entities (e.g. automatic doors predict the empty space they are calibrated to, opening when a person appears and violates the prediction). However, as systems, non-biological entities generally change in predictable ways over time, and, once learned, their input–output relationships often remain fixed. Thus, for some simple non-biological entities (e.g. an automatic door), mental inference (i.e. narrowing the hypothesis space with prior knowledge, then exploring it behaviorally through trial and error) may only have to occur *once*, after which the input–output relationship is known. That is, the internal model (X_i) of simple non-biological entities can be learned; and once learned, it generally does not change. A complete exploration of this line of thought is beyond this paper, but this framework may help distinguish the "intentional" and "mechanical" stances proposed by ([73]; for review, see [283]). An intentional stance may involve the full iterative process of first applying prior knowledge to limit the hypothesis space, then exploring it via interactive inference—a process that typically arrives at an approximate mental inference, at best. By contrast, for non-biological entities, the same process of inference can be applied, but generally, it only has to be applied a few times, or sometimes only *once*. After the simple non-biological model has been inferred, a mechanical stance can be applied. For example, the input–output relationships within a clock do not change, rendering it unnecessary to reapply the expensive, iterative process of interactive inference (unless the clock breaks, in which case it is anthropomorphized; [297]).

[17] In more computational terms, situational and dispositional inferences act as inductive biases, whereas interactive inference capitalizes on variance, explores the hypothesis space, and feeds into the priors for inductive bias [135].

hypothesis can be formed using prior knowledge, but to generate a more precise inference about others' expectations she must implement a behavior, experience prediction error, and refine her hypothesis. This has obvious implications for applications of machine-learning to mental inference (e.g. training algorithms to recognize emotions), as such approaches receive data, but do not generally interact with the humans that the data came from. In other words, human proficiency in mental inference may stem less from our ability to infer latent mental states from detached observations, and more from our ability to engage in a dynamic process of discovery, refining mental inferences by engaging with others and interpreting their reactions.

Second, this account of mental inference may provide a foundation to explain more complex phenomenon, such as social cohesion (i.e. group-formation; [231]). Prediction error has a metabolic cost (section 2.3) and affective consequences (i.e. arousal; section 3.2), meaning that mental inference (and especially the iterative process of interactive inference) may be more metabolically costly or more affectively aversive when dealing with unfamiliar others. For example, if Amelia has very little prior experience with a person (or a group of people) then her initial hypotheses are more likely to be inaccurate, prolonging the process of interactive inference and increasing its metabolic costs. In social interactions with unfamiliar others, Amelia may consistently form incorrect hypotheses and experience the interaction as stressful (i.e. when she and her partner do not share a language or social background, conversation may be halting and awkward). In some of these cases (e.g. when dealing with an outgroup member), Amelia may opt to avoid social interaction entirely, selecting a less expensive behavior, such as avoiding or excluding the unfamiliar other—that is, she may "choose a response that will most rapidly and completely reduce . . . her arousal and that incurs the fewest net costs" [75, p. 383]. Conversely, social interactions where both Amelia and her partner form accurate hypotheses about each other may "feel right": she and her partner have both converged on the expectations of the other, interacting in a way that minimizes arousal, prediction error, and metabolic costs ([43]; see also, [228, section V], for an illustrative example of affect guiding interpersonal behavior).

Although the present paper does not allow for more than a brief sketch of this account of social cohesion, our approach to modeling mental inference coheres nicely with a recent account of the relationship between social cohesion and processing fluency [231]. This account aimed to explain how social interaction recursively reinforces social cohesion: as people interact with each other, or as they coordinate their behavior around shared parts of the environment (e.g. ritual, dance; see also, [111]), their behavior becomes mutually fluent—i.e. it produces a mutual feeling of fluid and easy cognitive processing [232], stemming from the easy mutual exchange of information between two or more people [231]. This interpersonal fluency leads to mutual liking (e.g. [43,303]), which leads to even more behavioral coordination (e.g. [277]), which leads to yet more interpersonal fluency (e.g. [1]), producing a positive feedback loop that is thought to facilitate the creation of a cohesive social group. Further, if people (e.g. members of a mutual ingroup) begin with "similar attitudes, behaviors, or modes of communication . . . [then] interpersonal predictability increases", meaning that "similarity breeds liking partly because similarity increases interpersonal fluency" [231, p. 54]. In the language of our present work: fluent social interactions are metabolically efficient social interactions, in which the use of interactive inference—a metabolically expensive, trial-and-error-based process—can be minimized. When dealing with similar others (e.g. "in-group" members) Amelia can draw on more prior knowledge to accurately predict others' expectations and conform to them, minimizing her prediction error (and its affective consequences) during the dynamic process of mental inference.

3.5. Social influence and a sense of should

We began section 3 by proposing that individual humans (e.g. Amelia) learn to use a non-coercive form of social influence to control the behavior of others; specifically, Amelia conforms to others' expectations in order to regulate predictability in her social environment. Here, we briefly clarify what we mean by "social influence", distinguishing the non-coercive influence produced by conformity from social influence aimed at bringing about specific behaviors in others.

Social influence refers to behaviors enacted to affect the behavior of other people in a desired way. To distinguish the influence exercised by conformity from influence that produces specific behaviors in others, we coin two terms: *stabilizing influence*, and *directing influence*. *Stabilizing influence* is the control exercised by Amelia over others when Amelia *conforms to their expectations*. That is, when motivated by a *sense of should*, Amelia attempts to downregulate changes in others' behavior by conforming to their expectations, which in turn down-regulates prediction error in her social environment (and its metabolic costs). Throughout this paper, we have emphasized that conform-

ing to others' expectations is individually advantageous. *Stabilizing influence*, however, does not benefit Amelia by making others perform specific behaviors; rather, it benefits her by making others less likely to react to Amelia in unpredictable ways. To direct others to enact specific behaviors, *directing influence* would be necessary. For example, Amelia could make sounds (e.g. words) or move her body (e.g. point) in ways that cause other people to perform desirable actions (e.g. passing the table salt; see section 4.2.2). *Directing influence* is used as an umbrella term to encompass a variety of strategies, such as coercion by physical force or its threat, the shaping of behavior and beliefs by teaching (e.g. [192,311]), and normative influence (which is an umbrella term itself, encompassing behavioral change guided by reputation-seeking or by a *sense of should* in the person being influenced; see Fig. 1).

In some cases, *stabilizing influence* and *directing influence* may complement each other. If Amelia, motivated by a *sense of should*, regulates her social environment by conforming to others' expectations, then *directing influence* could be exercised over her simply by alerting her to those expectations, giving her an opportunity to exercise *stabilizing influence* of their own accord. In other words, *the most straightforward way to control someone else's behavior may be just to make them aware of what you want*. This can be made even more concrete with an example. Mameli [192] gives the following:

> "A father expects his children to share his own values. The father's expectations put a lot of psychological pressure on the children. As a result of this, the children end up valuing, at least in part, the same things as their father." [192, p. 609]

How does "psychological pressure" work in this example? In some cases, the father's *directing influence* may not be complemented by *stabilizing influence* and a *sense of should* in his children. For example, the father's *directing influence* might operate via reputation-seeking (Fig. 1), where the children might conform to "gain or maintain [his] acceptance" [167, p. 411], to avoid "social sanctions" from him [46, p. 1015], to achieve "social success" [217, p. 556], or "signal belongingness to a group [e.g. the family]" [285, p. 580]. In some cases, this reputation-seeking explanation could be correct. For example, a teenager might fear being shunned by his family if he were to come out as gay (or alternatively, he may curry favor to secure an inheritance). However, in other cases, the father's *directing influence* may be facilitated by the *stabilizing influence* his children exercise. That is, simply by knowing what their father expects, children will be motivated by a *sense of should* to conform to his expectations, a behavioral strategy that exercises control over their social environment, makes it predictable, and minimizes the metabolic costs and aversive affect it might otherwise impose. In the same way, a warden might regulate the behavior of prisoners simply by placing them under observation [94].

Directing influence via reputation-seeking and *directing influence* via a *sense of should* (i.e. *directing influence* complemented by *stabilizing influence*) both affect behavior, and the distinction being drawn between them is subtle. Astute readers will probably have noticed that the children (or prisoners) are still avoiding a cost in both pathways of *directing influence* (i.e. avoiding reputational sanctions in one case, and avoiding the metabolic costs of social disruption in another). However, when the children exercise *stabilizing influence*, it was not necessary for the father to make any explicit threat of sanction (see section 4.4 for further implications of this point). Instead, he simply made his expectations known, which provides his children with the knowledge that they need to self-regulate their behavior and produce an affectively and metabolically desirable social environment. Also, note that this pathway of *directing influence* (operating via a recipient's *sense of should* and *stabilizing influence*) only works when the recipient is both capable of understanding what is expected of him (i.e. he understands the message sent by the influencer) and is willing and able to exercise *stabilizing influence*. Of course, these prerequisites are not all or nothing, and can be satisfied by degree (we discuss some limited forms of *stabilizing influence* used by pigtail macaques in section 4.1.2), meaning that this *sense of should* mediated pathway of *directing influence* may vary in efficacy across individuals and across development.

A *sense of should*, then, allows individuals to exercise *stabilizing influence*, indirectly controlling the behavior of others by enacting the particular behaviors that they expect. It has been hypothesized that human prosocial behavior is made possible by our ability to regulate the behavior of others (i.e. by engaging in "mindshaping"; [192,196,197, 310,311]), and what we have suggested is subtly different (but highly complementary, see section 4.2.2): *we suggest that humans regulate the metabolic costs of their social environment by allowing their own behavior (and possibly their beliefs; see section 4.5) to be shaped by the expectations of others*. Allowing oneself to be guided by others' expectations almost certainly comes with advantages and disadvantages: it may benefit individuals by allowing them

to collectively stabilize dense social environments (e.g. an airplane packed full of animals), but it may also leave individuals open to manipulation by others, or even calcify social orders (e.g. hierarchies). Nonetheless, we suggest that influence is a two-way street, and that individuals that conform should not be understood as passive recipients of social pressure, but rather, they should be understood as active agents, engaged in an individually advantageous strategy of social-environmental regulation.

3.6. Summary

Conforming to others' expectations optimizes a predictable social environment. This predictability allows people to *coast* on (i.e. exploit) the accurate, metabolically efficient predictions made by their internal model about incoming sensory signals from the world (section 2.3.1); and further, such predictability could facilitate long-term social planning, either for cooperative or competitive ends. Consciously, this motivation to conform is experienced as a *sense of should*. A *sense of should* is separable from a desire for reward and an aversion to punishment ([7, chapter 12]; [268, III, 2.7]) and stems from the anticipation and interpretation of the arousal that occurs when processing prediction error. Developmentally, a *sense of should* likely emerges from the gradual accumulation of experience in domain-general processes, such as memory and associative learning; but critically, to know what behavior will satisfy a *sense of should*, children must develop an ability to infer the expectations of others—i.e. an ability to engage in mental inference. In conjunction with prior knowledge (e.g. dispositional and situational inferences) and potentially other proficiencies and cognitive processes [4,116,247], we hypothesize that mental inference is supported by a process of *interactive inference*. That is, by violating others' expectations (i.e. exploring via controlled "experiments") people can infer what behaviors were expected by others, and what behaviors were not. These inferences can be used to *construct* a more accurate model of the social environment. Conforming, then, indirectly regulates the behavior of others, and should be recognized as a form of social influence in and of itself (i.e. a *stabilizing influence*), and we suggest that learning how one's own behavior affects the predictability of others is almost certainly critical for maintaining one's own metabolically efficient existence within a society.

4. Extensions and implications of a *sense of should*

In this section, we highlight the implications of our account. First, we briefly clarify some common misconceptions about our framework. Second, we demonstrate its explanatory scope, exploring its relation to (what appear to be) two disparate phenomena: status quo biases, and social communication. Third, we elaborate on the relation between a *sense of should* and existing work in behavioral economics and game-theory, which has traditionally focused on motives related to reputation-seeking and material rewards, i.e. Adam Smith's first motive [268, III, 2.6–2.7]. Fourth, we highlight that expectations can motivate specific behaviors *even when the content of that expectation is evolutionarily irrelevant*, an implication that provides a concrete mechanism for the propagation of culture, and complicates nativist and functional accounts of behavior (e.g. [56]). Finally, we make clear that a *sense of should* may apply not only to behavior, but to beliefs as well, suggesting how social influence may affect the beliefs we adopt and maintain.

4.1. Complications and clarifications

The framework developed in section 3 suggests that conforming to others' expectations controls the metabolic costs of prediction error in a social environment, and that the metabolic costs (and affective experience) of social prediction error may condition individuals to prospectively avoid violating others' expectations. Some points of confusion may remain, and we briefly cover the most common ones here, while also drawing attention to some subtleties of the approach.

4.1.1. Precision terms and expecting the unexpected

In our approach, individuals regulate the metabolic costs of prediction error by inferring others' expectations and conforming to them. However, people sometimes want to be surprised (e.g. by gifts). Our approach can easily accommodate the observation that people like giving and receiving gifts, as well as the observation that even when giving a surprise gift there are limits to how surprising it can be.

First, although we have focused on explaining a *sense of should*, we do not suggest that people are *only* motivated to conform, nor that people are only motivated to minimize prediction error and metabolic costs (see section 2.3.1). When you surprise someone with a gift, the resultant prediction error provides information that helps *construct* a better model of the world (e.g. you learn how to buy better gifts in the future). This epistemic motivation must coexist with a motivation to regulate metabolic costs (see [103]). Further, prediction error generates arousal, and arousal is not necessarily valenced (e.g. [24,239,306]). As discussed in section 3.2, arousal can be experienced as positive [182], meaning that prediction error can be pleasurable in some contexts (a full account of how context guides the interpretation of arousal is beyond our present scope).

Second, even in gift giving, there are limits to how much others' expectations can be violated (e.g. some gifts are inappropriate). To account for this, precision terms (which, for simplicity, we have omitted in this paper) would need to be added to our model. Precision terms describe the precision with which a prediction is issued (e.g. [83]). Predictions can be precise and easily violated, or imprecise, in which case a range of sensory experiences can count as a "correct" prediction. When someone "expects the unexpected", as when receiving a gift, precision terms are likely adjusted to allow for some predictions to be violated and not others. For example, when someone expects to be surprised they may predict the arousal they are going to experience, but issue less precise predictions about the sensory signals that will trigger that arousal (i.e. what the gift will look like). Gift-giving can be conceptualized as trying to give someone the arousal he expects. That is, even for a surprise gift, he most likely has *some* expectation of the arousal he will experience, and if features of the gift evoke substantially more or less arousal than predicted (e.g. it is unreasonably expensive, or conversely, boring) then his arousal response may fall outside the predicted range. In this case, we hypothesize that his internal model will integrate the prediction error (i.e. information), and his behavior is more likely to change, consistent with our model (Equation (4); Fig. 2).

4.1.2. Is a sense of should unique to humans?

The framework outlined in section 3 does not depend on any specialized or uniquely human adaptations or mental processes. Given this, one might ask what makes humans unique, and why social pressure and norms (i.e. collective expectations) do not appear to motivate behavior in non-human primates. Our answer is that non-human primates do organize into social structures that are consistent with what our framework describes ([90,92,91]; also, see [160]), although with much less sophistication than humans. Humans sociality may not be categorically distinct from social behavior in other organisms; rather, complex human sociality may emerge when a threshold in domain-general cognitive abilities (e.g. memory) makes new strategies metabolically efficient.

A colony of pigtail macaques was observed, and all interactions among individuals were recorded [90,92,91]. Macaques would occasionally fight for resources or dominance, and wins and losses in these fights provided each macaque with cumulative information about their fighting ability relative to other individuals. When prior interactions made it clear that one macaque would likely lose to another, the lower-ranking macaque would bare its teeth and signal subordination to the dominant individual [91]. This subordination display has been suggested to act as a primitive social contract, which reduces the costs of social interaction for both the dominant, and the subordinate macaques [90]. Of course, subordination comes at a cost, and as a condition of this social contract the subordinate macaque must yield resources that interest the dominant individual [91]. But, by yielding, the subordinate macaque can keep its social environment predictable and avoid the risks of engaging in a fight.

In our framework, individuals, like these macaques, can exercise *stabilizing influence* (section 3.5), conforming to others' predictions to maintain a predictable social environment. To do this proficiently, however, people must infer others' predictions through mental inference (section 3.4). Compared to other primates, humans are exceptionally capable of mental inference [40,76] and we hypothesize that this ability stems from domain-general improvements (e.g. in memory, associative learning). For example, improvements in memory may allow more prior experience to be drawn upon to generate predictions about behavior (e.g. via dispositional and situational inferences). It may even be that a tipping point exists where it becomes more metabolically efficient to "keep the peace"—i.e. when mental inference becomes sufficiently precise, more metabolic efficiency may be gained (on average) by inferring and conforming to others' expectations than would be lost by forgoing the self-interested pursuit of reward. However, conforming to others' expectations precisely will be difficult when an animal lacks the cognitive capacity (or experience with another individual, group, or species) that is necessary to make precise mental inferences in the first place.

4.2. Extensions to known phenomena

As a general framework for understanding social motivation, a *sense of should* can point to dynamics underlying well-known social phenomena. Here, we provide an example of two phenomena that would ordinarily appear distinct. First, a *sense of should* may explain the dynamics of a status quo bias: why it manifests, why it is maintained, and when it may be overcome. Second, a *sense of should* may explain the adaptive advantage granted by communication and language.

4.2.1. Extension to a status quo bias

Why do people accept unjust institutions, and when do they resist? Although prior accounts of institutions could not provide a unified answer to this question [257, p. 108], our account may offer one. If a *sense of should* is widespread in a population, then all individuals contribute to (and benefit from) a shared social environment. That is, each individual regulates the predictability (and metabolic costs) of her own social environment by conforming to others' expectations; but because *all others* are doing the same (also for their own metabolic benefit), each person makes a small personal sacrifice but gains from (and contributes to) the increased predictability of the shared social environment. A self-interested strategy inadvertently benefits others. Given the collective metabolic benefits of social predictability (and the negative affective experience of social unpredictability), everyone has some motivation to maintain the social order—i.e. to maintain a status quo [163,244].

If the widespread adoption of a *sense of should* makes the social environment more predictable, then conversely, it also increases the potential costs (to each individual) of disrupting the status quo. The collective benefits are fragile: one non-conformist could disrupt the social environment for everyone. This may deter free riders, as self-interested actions that violate others' expectations also disrupt the free rider's social environment. However, this exact same dynamic may contribute to the oppression of minority groups. Any individual must weigh the costs and benefits of conformity: if she conforms then the social environment remains relatively predictable, and if she resists then things may get better, but they may get worse too. For example, she could call attention to a sexist comment—gambling on whether she will find support or face a backlash from the broader community—or she can let it slide, absorbing the offense and leaving the social environment unperturbed. The more predictable the current social environment, the more she has to lose by disrupting it (see, [70]). This line of reasoning suggests, unfortunately, that people may maintain, or even defend, an oppressive (but predictable) status quo, *so long as they are not suffering intolerably, and so long as the social arrangement is perceived as stable*.

But when the social arrangement is no longer perceived as stable, things may change quickly. If maintaining the status quo does not grant dividends in predictability (or if other costs outweigh the benefits of predictability), then there is no longer reason to maintain it. Consistent with this, Martin Luther King Jr. described the experience of being Black in the American South as being "harried by day and haunted by night by the fact that you are a Negro, living constantly at tiptoe stance, never quite knowing what to expect next" [171]. Oppressed minorities, then, may be uniquely positioned to challenge and change a status quo ([207,206]; for review, see [305]).

Such changes, however, may be opposed by some—specifically, by those for whom the *status quo is still bearable*. Even among people who consider themselves supportive of the oppressed, there will be some "who [are] more devoted to 'order' than to justice; who [prefer] a negative peace which is the absence of tension to a positive peace which is the presence of justice; who constantly [say]: 'I agree with you in the goal you seek, but I cannot agree with your methods of direct action'" [171]. Social change is always weighed against the alternative: doing nothing. Given this, minorities may be more likely to win concessions when the status quo is made unsustainable for the majority as well—that is, when protest creates "a situation so crisis packed that it will inevitably open the door to negotiation" [171]. In other words, if the option of maintaining a predictable status quo is ruled out, then people may be forced to look for and implement solutions.

4.2.2. Extension to communication and language

Communication may seem far afield from a *sense of should*, but it emerges within our approach as another means of regulating the social environment. As discussed throughout this paper, social environments can be made more predictable by conforming, i.e. by choosing a behavior, $\boldsymbol{b}$, that minimizes $|\boldsymbol{b} - \boldsymbol{pb_i}|$ in Equation (4):

$$pe_i^{Ext:R} \propto |\boldsymbol{b} - \boldsymbol{pb_i}|$$

where

$pe_i^{Ext:R}$ represents your reciprocal prediction error, from one entity (i) in the environment,

b represents your behavior,

pb_i represents the prediction of one entity (i) about your behavior.

Communication takes the opposite approach. Rather than changing your behavior to conform to others' expectations, communication involves guiding others to more accurately predict your behavior; specifically, by issuing sensory signals. That is, mutually understood sensory signals (e.g. sounds forming words) may affect the predictions and behavior of others (see speech act theory; [10]). By issuing signals that affect others' predictions, pb_i, those predictions may be made to more closely match your behavior, b, minimizing $|b - pb_i|$ at some future moment. For example, when borrowing a colleague's pen, rather than snatching it from her desk, you might use a declarative claim to signal your intent by saying: "I'll use this pen" (or, intention can be signaled more subtly with the interrogative: "Can I borrow your pen"?; [242]). By guiding your partner's predictions, you may reduce the prediction error they experience when you do reach across the table.[18] In this context, you personally benefit by helping others make accurate predictions about your behavior[19]—the more accurately others can predict your behavior, the less need there is for you to infer their expectations and conform. This strategy of guiding others' inferences is consistent with "mindshaping" proposals [192,196,197,310,311], which suggest that an ability to guide the mental inferences of others may precede mental inference. Our framework suggests that mindshaping (to guide others' expectations) and mental inference (to infer and conform to others' expectations) are complementary, and each one likely bootstraps improvement in the other. Thus, the same contingency between your behavior and others' predictions that gave rise to a *sense of should* also facilitates communication, where mutually understood sounds (i.e. words) affect the predictions that others make.

Communicating to signal intent, as in the example above, would be adaptive even if no one else were motivated by a *sense of should*—it does not require that others feel any motivation to conform to your expectations; rather, it only requires that their brains encode information as prediction error. But, as briefly discussed in section 3.5, if others do experience a *sense of should* then communication can be even more advantageous: ordering others (via an imperative; [242]) *can communicate your expectation, and if others know what you expect of them, then their behavior can be motivated by a sense of should*. In other words, if other people exercise control over their social environment by conforming to your expectations (i.e. they exercise *stabilizing influence*; section 3.5), then you can directly affect their behavior by making your expectations known (i.e. you can exercise *directing influence*). For example, if you say "please pass the salt", it would be disruptive for someone to refuse without good reason (see [2,15]), or without offering an excuse [286]. Thus, by communicating your *intention* to behave in certain ways (e.g. via declarative claims) you may guide others (via mindshaping) to correctly predict your behavior, but by communicating your *expectations* (e.g. via imperatives) you may exercise *directing influence* over the behavior of others (so long as they are motivated by a *sense of should*).

It has been suggested that ultimate goal of communication should be understood as influencing others' conduct (i.e. behavior), as "with any reasonably broad definition of conduct, it is clear that communication either affects conduct or is without any discernible and probable effect at all" [266, p. 5]. Language is well beyond the scope of this paper, but future research could explore how these individual benefits of communication—i.e. making your own behavior predictable, and influencing the behavior of others via a *sense of should*—interact with the development of language, elaborating both on how each person's internal model (X_i in Equation (4); Fig. 2) mediates the interpretation

---

[18] One might object that prediction error has not been reduced, it has only been moved forward in time—i.e. your partner receives prediction error from the spoken words, rather than from your reaching across the table. However, the magnitude of prediction error from your words and from your reaching are not the same. The reason they are not the same can be explained by the precision of predictions (which were omitted from model development in section 3, but discussed in section 4.1.1). When you speak, your colleague does predict that you will make sounds (which are somewhat low-precision, as she doesn't know exactly you will say); however, she has no reason to predict that you will make a sudden physical movement toward her. By communicating your movements in advance, you are issuing a predictable signal (i.e. sounds) to change her predictions in another sensory domain (i.e. her predictions about your movements).

[19] Of course, this may create a niche for counter-strategies, such as deception (as discussed in section 3.1). If most people communicate accurately—helping others accurately anticipate their behavior—then others may lie, leveraging the expectation of honesty for their own advantage. But for deception to work there must be a general assumption that others are truthful [164]. Indeed, a fundamental principle of language is to "try to make your contribution one that is true" [134, p. 27]. Language may provide a powerful and general tool for directly affecting others' predictions, meaning, conversely, it can be powerfully abused.

of symbols (e.g. spoken and written word) into their intended meaning, and on how that intended meaning affects behavior [10].

4.3. Implications for behavioral economics and game theory

Assumptions about what motivates human behavior are foundational to economic theory, and our account describes a novel route through which others' expectations motivate behavior. Although many economists and game-theorists are agnostic about *why* humans behave as they do [237,243], others offer evolutionary accounts of human behavior (e.g. [82,183,211,288]). In these accounts, motivation is typically characterized as a desire (perhaps an unconscious desire; [294]) to maximize evolutionary fitness in terms of reputation or material reward. If people behave altruistically, then it is assumed that the possibility of reciprocity or the threat of third-party punishment motivates the behavior (for review, see [183]). Our account (following [268]) does not rule out the benefits of these approaches (as behavior can be multiply determined; Fig. 1), but it also focuses on a different motivation: a *sense of should*. A *sense of should* could be applied as an alternative explanation for a range of topics (e.g. the motivational force of observability; avoiding others who would ask for help, but helping readily when asked; cooperating when others are also predicted to cooperate; see [229]), but due to space constraints, we focus on one example: the individual benefits of a *sense of should* within a prisoner's dilemma game.

A prisoner's dilemma game involves two players, each of whom chooses to either cooperate with or betray their partner. If both cooperate, then both win a modest payoff (e.g. $3). If one player chooses to cooperate and is betrayed, then he receives nothing (the sucker payoff). Betraying a partner gives the maximum payoff if the partner cooperated (e.g. $5) and gives a meager payoff if the partner also chose betrayal (e.g. $1). The rational choice, if the game is played only once, is to choose betrayal. If the game is played iteratively, however, then individual benefits are maximized by cooperating [11,12,288]. In the context of a prisoner's dilemma game, selfishness is a short-term strategy, whereas cooperation provides long-term benefits [229].

Conforming to others' expectations (via a *sense of should*) is advantageous because it minimizes the metabolic costs of an unpredictable social environment. However, the context of a prisoner's dilemma game already involves a massive reduction in uncertainty, which in turn changes what strategies are applicable. In a prisoner's dilemma, the only unknown factor is the other player's choice, and the reward for each combination of moves is known. However, in the real world, many contingencies are unknown. Favors may or may not be reciprocated. Violations may or may not be punished (indeed, third-party punishment is rare outside of economic games; [179,219,220]). If rewards and punishments often fail to materialize as expected in the real-world, then in the real-world both selfishness and reciprocity are *risky* strategies. By contrast, a *sense of should* is a *safe default*: it optimizes a predictable social environment and in doing so produces a small, but reliable and immediate metabolic reward (akin to the "safe" option in a delayed-discounting task; [173]). By conforming, the real-world social environment is made more predictable, social information processing is made more efficient, and long-term social planning is made possible (section 3.1). With respect to altruistic behavior (i.e. conforming to others' expectations that you will help), the small costs of helping others need not be repaid by reciprocity, as a predictable social environment is rewarding in itself (and an unpredictable environment would be metabolically disadvantageous). However, if a *sense of should* makes the social environment predictable, then it is less useful in environments that are already tightly controlled. Prisoner's dilemma games rule out the benefits of a *sense of should* by design: their controlled structure eliminates the conditions where a *sense of should* is most adaptive.

4.3.1. Relation to psychological game theory and guilt aversion

Behavior is multiply determined, and emerges from competing motivations (e.g. monetary reward vs. a *sense of should*), and some cleverly designed economic games have captured this insight, particularly in the domain of psychological game theory and guilt aversion ([28,29,114,42]; also, see [3]). In this research, the beliefs and emotions of the players are considered directly relevant for modeling behavior. For example, Chang and colleagues [42] modeled guilt in a one-shot trust game, where an investor can give money to a trustee; if they do, then the investment is multiplied and the trustee can return any amount. As the trustee can keep everything without reprisal, it follows rationally that the trustee should return nothing, and therefore, the investor should invest nothing. Contrary to this logic, the trustee (player 2) generally returned what he thought the investor (player 1) expected back. To capture this,

overall utility for the trustee was modeled as a function of the money he received, minus his guilt. Guilt was modeled as:

$$\Theta_{12}(E_2 E_1 S_2 - S_2)$$

where

Θ_{12} is a guilt sensitivity parameter, modeling whether the trustee (player 2) cares about violating the investor's (i.e. player 1's) expectation,

$E_2 E_1 S_2$ represents the amount that the trustee (player 2) believes the investor (player 1) expects, and

S_2 represents the amount that the trustee actually returns.

Thus, in the guilt aversion model, the trustee is motivated to choose a behavior that matches the investor's expectation.

Analogously, in our model, Amelia is motivated to choose a behavior that conforms to Bob's expectation. According-ing to equation (5):

$$pe_i^{Ext:R} \propto |b - (pb_i^M + e)|$$

where

pb_i^M represents Amelia's estimate (i.e. her mental inference) of Bob's prediction about her behavior,

e represents the inaccuracy in Amelia's estimate, such that $pb_i^M + e = pb_i$, and

b represents Amelia's actual behavior.

These models are equivalent, where

$$(pb_i^M + e) = E_2 E_1 S_2$$
$$b = S_2$$

Starting from drastically different foundations, we reached the same conclusion as Chang and colleagues: individuals are motivated to minimize the discrepancy between their behavior and the inferred expectations of others.

A *sense of should* and guilt aversion only diverge in their accounts of what emotions are and how they work. Guilt aversion is defined as a motivation to avoid the aversive consequences of "failing to live up to others' expectations" ([30]; quoted in [28, p. 170]). Our approach to modeling a *sense of should* leverages a more general understanding of emotions as constructed explanations for allostatic changes, behaviors, and their associated sensory consequences, in-cluding affect [22,23]. In the theory of constructed emotion, "guilt" is a word that refers to a *category* of heterogeneous instances, each instance tailored to a specific context or situation [19,22,23,189]. A category is a group of instances that are similar in some way. In the context of guilt, that similarity is provided by a *sense of should*: the features of each instance of guilt (e.g. the physiological changes, the behavior performed, the affect felt) vary according to the requirements of the situation, but all of this variation is united by the motivation to conform to other's expectations. It would be a mistake, however, to conclude that a *sense of should* is synonymous with guilt. Guilt is a specific example under the umbrella of a *sense of should*, whereas a *sense of should* is a more general motivation to conform to others' expectations, and as such, it can be observed in the context of other emotional instances (e.g. a servile lackey doing whatever his boss asks; a parent caving to his demanding child).

4.4. Implications for culture, norms, and evolutionary psychology

In our framework, individuals benefit from conforming to others' expectations, thereby optimizing a predictable social environment. The logic works regardless of the content of those expectations. It doesn't matter what, specif-ically, others expect you to do: to regulate social predictability it only matters that your behavior matches others' expectations, whatever they may be. This is a powerful implication, as it provides a general avenue through which culture (i.e. the collective or common expectations of others) can motivate individuals to adopt ways of behaving [84, 118,145,147,234]—and perhaps even ways of thinking [149,150]—orthogonal to whether the content of the behavior serves any functional end. Foundational evolutionary models have shown that "punishment allows the evolution of cooperation (or anything else) in sizable groups" [35, p. 171]; likewise, any arbitrary behavior could be motivated by a *sense of should*, and no explicit or costly punishment is even required to reinforce it. As discussed in section 3.2, a *sense of should* is a punishment that the brain "inflicts on itself" (p. 34), and it is a punishment that does not need to be intentionally administered by anyone—it only requires that others behave less predictably when their expectations are violated.

If expectations motivate behavior, then these expectations may begin as historical accidents, but, as they propagate across generations [71,143], they may become cemented into the foundation of social reality (i.e. expectations shared across many people). For example, it has been hypothesized that elements of western individualism [146] trace their origins to the Marriage and Family Program of the Catholic Church [250]. This program promoted 'by choice' marriages, required couples to set up independent households, and forbid marriage between immediate cousins (later extending the prohibition to distant cousins, step-siblings, and in-laws). Why the program was implemented remains debated, but hypotheses are firmly rooted in the dynamics of medieval politics—for example, the policy disrupted the inheritance of property within clans, which freed individuals to give up their property to the church. There is reason to believe, then, that the norms governing us today (e.g. regarding partnership, familial impendence, and even incest) are completely infused with the flotsam of history. A *sense of should* points to where the work of psychologists, historians, and anthropologists might intersect, and makes clear how historical accidents and social constructs continue to affect behavior today.

This hypothesis, that one person's behavior can be motivated by others' expectations—*and not necessarily* by any advantage granted by the behavior itself (as, of course, many cultural innovations are adaptive; [145])—poses serious problems for hypotheses about innate, functionally adaptive cognitive modules and intuitions (e.g. [57,138,139,287]). You were born into a world where other humans already have expectations about how you will behave: they expect, for example, that you will respect ownership, exchange money for food and treat men and women differently (for a review of expectancy effects in gender development, see [192]). It is possible, of course, that evolved functional modules motivate particular behaviors (or that existing expectations trace their origins to innate features of human cognition[20]), but in practice, for the purposes of motivation via a *sense of should*, it doesn't matter where an expectation came from. To satisfy a *sense of should*, only current expectations matter; their origin does not. Explanations of behavior, rooted in a *sense of should*, then, represent a potent alternative to the functional view, and in the absence of direct evidence for the functional evolutionary origin of any particular behavior, this is an alternative that cannot be dismissed.

4.5. Implications for adopting and maintaining beliefs

Thus far, we have discussed a *sense of should* as motivating behavior. But if others are capable of precise mental inference, then the motivation may extend beyond behavior: others' expectations may compel you to adopt or maintain *beliefs*. For example, imagine having the faintest thought that you no longer love your partner of 20 years. Your behavior around your partner might change, even subtly. Your partner, knowing you well (i.e. having many prior experiences to draw on for mental inference; section 3.4), might notice that something is strange and attempt to infer the cause. *The mere presence of your belief, then, may increase the likelihood of social disruption*, and if holding the belief is threatening then you might feel a pressure to dispel it [132]. The expectations of close social relations (i.e. those most capable of accurate mental inference, and most central to your social environment; [89]) may shape your behaviors and beliefs. Extending traditional accounts of cognitive dissonance: you may not only be motivated to hold an internally consistent set of beliefs [88,87], but rather, you may be motivated to hold beliefs that maintain a predictable and metabolically efficient social environment. Some beliefs, then, may be formed or maintained through social influence, rather than through rational consideration of the evidence.

In the mid-20th century, this conclusion was seen as threatening to the entire enterprise of science [133]. It implies that a scientist's beliefs may not always have a rational origin—e.g. biologists may, in part, have come to believe in evolution because their colleagues do, rather than because they observed evidence that convinced them.[21] Instead, psychologists adopted theoretical perspectives informed by tenets of rationalism and individualism [133, chapter 6]. Both tenets have since been challenged—rationalism by accounts of affective motivation (e.g. [138]) and motivated cognition (e.g. [181]), and individualism by cross-cultural research highlighting the Western assumptions embedded in methods and theory [146]. Given the success of these perspectives, it should be trivial to accept that some beliefs

---

[20] But see [160] for another alternative: universal features of behavior (and even our moral commitments) could also arise from the combined dynamics of our biology and emergent social structures.

[21] Even objectively true beliefs (e.g. in evolution) could be acquired or maintained under the influence of social pressure. That is, beliefs are responsive to evidence, *but they are also responsive to the metabolic consequences of social disruption*. It may be difficult, at times, to untangle these two causes—e.g. as a child, you may learn, and firmly believe that the Earth orbits the sun, but never verify this with evidence (see also, [225]).

are not formed or maintained by impartial consideration of the evidence. More broadly though, our account implies that the shared expectations of scientific communities, like any other community, can exert a real influence on individuals and their interpretation of reality. Thomas Kuhn made a similar point [180]—that scientific communities establish theoretical frameworks, including shared sets of assumptions and expectations, that help scientists interpret and communicate their findings. *But*, these same communities and theoretical frameworks also create the conditions necessary for social influence via a *sense of should*. If one has built a scientific career, professional relationships, and a personal identity, that all depend on a particular theoretical framework—e.g. that current standards for statistical inference in psychology are acceptable (cf. [214]); that the brain is usefully considered as analogous to a computer (cf. [274]); or that discrete functions can be localized to brain regions (cf. [290])—then abandoning prior beliefs in the face of conflicting evidence will be socially and metabolically disruptive. This conclusion was threatening in the mid-20th century, and it may still be threatening now.

5. Conclusion

Asch claimed that psychology, with its focus on individuals, could potentially be to the social sciences what physics is to the natural sciences:

> "All great activities in society—economic, political, artistic—have their center in individuals . . . . And indeed, we find that the great social theorists, such as Hobbes, Rousseau, Adam Smith, and Marx, . . . in one way or another attempted to deduce, from a psychological starting point, consequences for political organization, economic practices, and education." [7, pp. 4–5]

We are psychologists, and so we have focused on the individual. We have attempted to deduce, from a biological starting point, the consequences of biological principles for individual social cognition—namely, how individuals are motivated by a *sense of should* to conform to the expectations of others. Although a *sense of should* is individually adaptive, it also creates the necessary conditions for communities, traditions, and eventually societies to take root and grow. Many individuals, then, all acting to optimize predictability for themselves, may collectively contribute to a common social reality: a predictable, socially constructed foundation on which societies can be built.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

References

[1] Adank P, Hagoort P, Bekkering H. Imitation improves language comprehension. Psychol Sci 2010;21(12):1903–9. https://doi.org/10.1177/0956797610389192.

[2] Andreoni J, Rao JM. The power of asking: how communication affects selfishness, empathy, and altruism. J Public Econ 2011;95(7–8):513–20. https://doi.org/10.1016/j.jpubeco.2010.12.008.

[3] Andrighetto G, Grieco D, Tummolini L. Perceived legitimacy of normative expectations motivates compliance with social norms when nobody is watching. Front Psychol 2015;6. https://doi.org/10.3389/fpsyg.2015.01413.

[4] Apperly IA. What is "theory of mind"? Concepts, cognitive processes and individual differences. Q J Exp Psychol 2012;65(5):825–39. https://doi.org/10.1080/17470218.2012.676055.

[5] Asch S. Effects of group pressure upon the modification and distortion of judgments. In: Guetzkow H, editor. Groups, leadership and men: research in human relations. Carnegie Press; 1951. p. 177–90.

[6] Asch S. Opinions and social pressure. Sci Am 1955;193(5):31–5. https://doi.org/10.1038/scientificamerican1155-31.

[7] Asch S. Social psychology. 7th ed. Prentice-Hall; 1962. (Original work published 1952).

[8] Attwell D, Laughlin SB. An energy budget for signaling in the grey matter of the brain. J Cereb Blood Flow Metab 2001;21(10):1133–45. https://doi.org/10.1097/00004647-200110000-00001.

[9] Atzil S, Gao W, Fradkin I, Barrett LF. Growing a social brain. Nat Hum Behav 2018;2(9):624–36. https://doi.org/10.1038/s41562-018-0384-6.

[10] Austin JL. How to do things with words. Oxford University Press; 1962.

[11] Axelrod R. The emergence of cooperation among egoists. Am Polit Sci Rev 1981;75(2):306–18. https://doi.org/10.2307/1961366.

[12] Axelrod R. An evolutionary approach to norms. Am Polit Sci Rev 1986;80(4):1095–111. https://doi.org/10.2307/1960858.

[13] Bach P, Schenke KC. Predictive social perception: towards a unifying framework from action observation to person knowledge. Soc Personal Psychol Compass 2017;11(7):e12312. https://doi.org/10.1111/spc3.12312.

[14] Baillargeon R, Scott RM, He Z. False-belief understanding in infants. Trends Cogn Sci 2010;14(3):110–8. https://doi.org/10.1016/j.tics.2009.12.006.

[15] Balafoutas L, Sutter M. On the nature of guilt aversion: insights from a new methodology in the dictator game. J Behav Exp Finance 2017;13:9–15. https://doi.org/10.1016/j.jbef.2016.12.001.

[16] Baldassano C, Chen J, Zadbood A, Pillow JW, Hasson U, Norman KA. Discovering event structure in continuous narrative perception and memory. Neuron 2017;95(3):709–21. https://doi.org/10.1016/j.neuron.2017.06.041.

[17] Barbas H. General cortical and special prefrontal connections: principles from structure to function. Annu Rev Neurosci 2015;38(1):269–89. https://doi.org/10.1146/annurev-neuro-071714-033936.

[18] Baron-Cohen S. Mindblindness: an essay on autism and theory of mind. MIT Press; 1997.

[19] Barrett LF. Solving the emotion paradox: categorization and the experience of emotion. Personal Soc Psychol Rev 2006;10(1):20–46. https://doi.org/10.1207/s15327957pspr1001_2.

[20] Barrett LF. Valence is a basic building block of emotional life. J Res Pers 2006;40(1):35–55. https://doi.org/10.1016/j.jrp.2005.08.006.

[21] Barrett LF. The conceptual act theory: a precis. Emot Rev 2014;6(4):292–7.

[22] Barrett LF. The theory of constructed emotion: an active inference account of interoception and categorization. Soc Cogn Affect Neurosci 2017;12(1):1–23. https://doi.org/10.1093/scan/nsw154.

[23] Barrett LF. How emotions are made: the secret life of the brain. Pan Macmillan; 2017.

[24] Barrett LF, Bliss-Moreau E. Affect as a psychological primitive. In: Advances in experimental social, vol. 41. Academic Press; 2009. p. 167–218.

[25] Barrett LF, Finlay BL. Concepts, goals and the control of survival-related behaviors. Curr Opin Behav Sci 2018;24:172–9. https://doi.org/10.1016/j.cobeha.2018.10.001.

[26] Barrett LF, Simmons WK. Interoceptive predictions in the brain. Nat Rev Neurosci 2015;16(7):419–29. https://doi.org/10.1038/nrn3950.

[27] Batson CD, Shaw LL. Evidence for altruism: toward a pluralism of prosocial motives. Psychol Inq 1991;2(2):107–22. https://doi.org/10.1207/s15327965pli0202_1.

[28] Battigalli P, Dufwenberg M. Guilt in games. Am Econ Rev 2007;97(2):170–6.

[29] Battigalli P, Dufwenberg M. Dynamic psychological games. J Econ Theory 2009;144(1):1–35. https://doi.org/10.1016/j.jet.2008.01.004.

[30] Baumeister RF, Stillwell AM, Heatherton TF. Guilt: an interpersonal approach. Psychol Bull 1994;115(2):243–67. https://doi.org/10.1037/0033-2909.115.2.243.

[31] Berns GS, Capra CM, Moore S, Noussair C. Neural mechanisms of the influence of popularity on adolescent ratings of music. NeuroImage 2010;49(3):2687–96. https://doi.org/10.1016/j.neuroimage.2009.10.070.

[32] Bicchieri C. The grammar of society: the nature and dynamics of social norms. Cambridge University Press; 2006.

[33] Blakemore S-J, Frith CD, Wolpert DM. Spatio-temporal prediction modulates the perception of self-produced stimuli. J Cogn Neurosci 1999;11(5):551–9. https://doi.org/10.1162/089892999563607.

[34] Blakemore S-J, Mills KL. Is adolescence a sensitive period for sociocultural processing?. Annu Rev Psychol 2014;65:187–207. https://doi.org/10.1146/annurev-psych-010213-115202.

[35] Boyd R, Richerson PJ. Punishment allows the evolution of cooperation (or anything else) in sizable groups. Ethol Sociobiol 1992;13:171–95.

[36] Braem S, Coenen E, Bombeke K, van Bochove ME, Notebaert W. Open your eyes for prediction errors. Cogn Affect Behav Neurosci 2015;15(2):374–80. https://doi.org/10.3758/s13415-014-0333-4.

[37] Brown JH, Gillooly JF, Allen AP, Savage VM, West GB. Toward a metabolic theory of ecology. Ecology 2004;85(7):1771–89. https://doi.org/10.1890/03-9000.

[38] Bullmore E, Sporns O. The economy of brain network organization. Nat Rev Neurosci 2012;13(5):336–49. https://doi.org/10.1038/nrn3214.

[39] Burghardt GM. The genesis of animal play: testing the limits. MIT Press; 2005.

[40] Call J, Tomasello M. Does the chimpanzee have a theory of mind? 30 years later. Trends Cogn Sci 2008;12(5):187–92. https://doi.org/10.1016/j.tics.2008.02.010.

[41] Chanes L, Barrett LF. Redefining the role of limbic areas in cortical processing. Trends Cogn Sci 2016;20(2):96–106. https://doi.org/10.1016/j.tics.2015.11.005.

[42] Chang LJ, Smith A, Dufwenberg M, Sanfey AG. Triangulating the neural, psychological, and economic bases of guilt aversion. Neuron 2011;70(3):560–72. https://doi.org/10.1016/j.neuron.2011.02.056.

[43] Chartrand TL, Bargh JA. The chameleon effect: the perception-behavior link and social interaction. J Pers Soc Psychol 1999;76(6):893–910.

[44] Christie ST, Schrater P. Cognitive cost as dynamic allocation of energetic resources. Front Neurosci 2015;9. https://doi.org/10.3389/fnins.2015.00289.

[45] Churchland P. Conscience: the origins of moral intuition. W. W. Norton, Incorporated; 2019.

[46] Cialdini RB, Reno RR, Kalgren CA. A focus theory of normative conduct: recycling the concept of norms to reduce littering in public places. J Pers Soc Psychol 1990;58(6):1015–26. https://doi.org/10.1037/0022-3514.58.6.1015.

[47] Clark A. Whatever next? Predictive brains, situated agents, and the future of cognitive science. Behav Brain Sci 2013;36(3):1–24. https://doi.org/10.1017/S0140525X12000477.

[48] Clark A. Surfing uncertainty: prediction, action, and the embodied mind. Oxford University Press; 2015.

[49] Clark M. Humour and incongruity. Philosophy 1970;45(171):20–32. https://doi.org/10.1017/S003181910000958X.

[50] Clarke DD, Sokoloff L. Circulation and energy metabolism in the brain. In: Basic neuro-chemistry: molecular, cellular and medical aspects. Lippincott-Raven; 1999. p. 637–70.

[51] Clark-Polner E, Wager TD, Satpute AB, Barrett B, Feldman L. Neural fingerprinting: meta-analysis, variation and the search for brain-based essences in the science of emotion. In: Barrett LF, Lewis M, Haviland-Jones JM, editors. The handbook of emotion. 4th ed. Guilford; 2016. p. 146–65.

[52] Claxton G. Why can't we tickle ourselves?. Percept Mot Skills 1975;41(1):335–8. https://doi.org/10.2466/pms.1975.41.1.335.

[53] Cohen JD, McClure SM, Yu AJ. Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. Philos Trans R Soc Lond B, Biol Sci 2007;362(1481):933–42. https://doi.org/10.1098/rstb.2007.2098.

[54] Conant RC, Ross Ashby W. Every good regulator of a system must be a model of that system. Int J Syst Sci 1970;1(2):89–97. https://doi.org/10.1080/00207727008920220.

[55] Constant A, Ramstead MJD, Veissière SPL, Friston K. Regimes of expectations: an active inference model of social conformity and human decision making. Front Psychol 2019;10. https://doi.org/10.3389/fpsyg.2019.00679.

[56] Cosmides L, Tooby J. Cognitive adaptations for social exchange. In: Barkow J, Cosmides L, Tooby J, editors. The adapted mind: evolutionary psychology and the generation of culture. Oxford University Press; 1992. p. 163–228.

[57] Cosmides L, Tooby J, Kurzban R. Perceptions of race. Trends Cogn Sci 2003;7(4):173–9. https://doi.org/10.1016/S1364-6613(03)00057-3.

[58] Craig AD. How do you feel? An interoceptive moment with your neurobiological self. Princeton University Press; 2015.

[59] Critchley HD, Tang J, Glaser D, Butterworth B, Dolan RJ. Anterior cingulate activity during error and autonomic response. NeuroImage 2005;27(4):885–95. https://doi.org/10.1016/j.neuroimage.2005.05.047.

[60] Crockett MJ. Models of morality. Trends Cogn Sci 2013;17(8):363–6. https://doi.org/10.1016/j.tics.2013.06.005.

[61] Crone EA, Somsen RJM, Beek BV, Molen MWVD. Heart rate and skin conductance analysis of antecendents and consequences of decision making. Psychophysiology 2004;41(4):531–40. https://doi.org/10.1111/j.1469-8986.2004.00197.x.

[62] Cushman F. Action, outcome, and value: a dual-system framework for morality. Personal Soc Psychol Rev 2013;17(3):273–92. https://doi.org/10.1177/1088868313495594.

[63] Damasio AR. The feeling of what happens: body and emotion in the making of consciousness. Houghton Mifflin Harcourt; 1999.

[64] Darwin C. On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life. In: Manis J, editor. Penn State University's Electronic Classics; 2001. (Original work published 1859).

[65] Daw ND, Gershman SJ, Seymour B, Dayan P, Dolan RJ. Model-based influences on humans' choices and striatal prediction errors. Neuron 2011;69(6):1204–15. https://doi.org/10.1016/j.neuron.2011.02.027.

[66] Daw ND, Niv Y, Dayan P. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nat Neurosci 2005;8(12):1704–11. https://doi.org/10.1038/nn1560.

[67] Dawkins R. The selfish gene: 40th anniversary edition. Oxford University Press; 2016. (Original work published 1976).

[68] Dayan P, Yu AJ. Phasic norepinephrine: a neural interrupt signal for unexpected events. Netw Comput Neural Syst 2006;17(4):335–50. https://doi.org/10.1080/09548980601004024.

[69] De Jaegher H, Di Paolo E, Gallagher S. Can social interaction constitute social cognition?. Trends Cogn Sci 2010;14(10):441–7. https://doi.org/10.1016/j.tics.2010.06.009.

[70] DeDeo S. Collective phenomena and non-finite state computation in a human social system. PLoS ONE 2013;8(10):e75818. https://doi.org/10.1371/journal.pone.0075818.

[71] DeDeo S. Major transitions in political order. In: Walker SI, Davies PCW, Ellis GFR, editors. From matter to life: information and causality. Cambridge University Press; 2017. p. 393–428.

[72] Denève S, Jardri R. Circular inference: mistaken belief, misplaced trust. Curr Opin Behav Sci 2016;11:40–8. https://doi.org/10.1016/j.cobeha.2016.04.001.

[73] Dennett DC. The intentional stance. MIT Press; 1987.

[74] Deutsch M, Gerard HB. A study of normative and informational social influences upon individual judgment. J Abnorm Soc Psychol 1955;51(3):629–36. https://doi.org/10.1037/h0046408.

[75] Dovidio JF. Helping behavior and altruism: an empirical and conceptual overview. In: Berkowitz L, editor. Advances in experimental social psychology, vol. 17. Academic Press; 1984. p. 361–427.

[76] Drayton LA, Santos LR. A decade of theory of mind research on Cayo Santiago: insights into rhesus macaque social cognition: rhesus macaque theory of mind. Am J Primatol 2016;78(1):106–16. https://doi.org/10.1002/ajp.22362.

[77] Dreyfus G, Thompson E. Asian perspectives: Indian theories of mind. In: Zelazo PD, Moscovitch M, Thompson E, editors. The Cambridge handbook of consciousness. Cambridge University Press; 2007. p. 89–114.

[78] Duranti A. Further reflections on reading other minds. Anthropol Q 2008;81(2):483–94. https://doi.org/10.1353/anq.0.0002.

[79] Edelman GM, Tononi G. A universe of consciousness: how matter becomes imagination. Basic Books; 2000.

[80] Emirbayer M. Manifesto for a relational sociology. Am J Sociol 1997;103(2):281–317. https://doi.org/10.1086/231209.

[81] Fee MS, Mitra PP, Kleinfeld D. Central versus peripheral determinants of patterned spike activity in rat vibrissa cortex during whisking. J Neurophysiol 1997;78(2):1144–9. https://doi.org/10.1152/jn.1997.78.2.1144.

[82] Fehr E, Gächter S. Altruistic punishment in humans. Nature 2002;415(6868):137–40. https://doi.org/10.1038/415137a.

[83] Feldman H, Friston KJ. Attention, uncertainty, and free-energy. Front Human Neurosci 2010;4. https://doi.org/10.3389/fnhum.2010.00215.

[84] Feldman MW, Laland KN. Gene-culture coevolutionary theory. Trends Ecol Evol 1996;11(11):453–7. https://doi.org/10.1016/0169-5347(96)10052-5.

[85] FeldmanHall O, Shenhav A. Resolving uncertainty in a social world. Nat Hum Behav 2019;3(5):426. https://doi.org/10.1038/s41562-019-0590-x.

[86] Felleman DJ, Van Essen DC. Distributed hierarchical processing in the primate cerebral cortex. Cereb Cortex 1991;1(1):1–47. https://doi.org/10.1093/cercor/1.1.1.

[87] Festinger L. A theory of cognitive dissonance. Stanford University Press; 1962.

[88] Festinger L. Cognitive dissonance. Sci Am 1962;207(4):93–106. JSTOR.

[89] Fiske AP, Rai TS. Virtuous violence. Cambridge University Press; 2014.

[90] Flack JC. Multiple time-scales and the developmental dynamics of social systems. Philos Trans R Soc Lond B, Biol Sci 2012;367(1597):1802–10. https://doi.org/10.1098/rstb.2011.0214.

[91] Flack JC, de Waal F. Context modulates signal meaning in primate communication. Proc Natl Acad Sci USA 2007;104(5):1581–6. https://doi.org/10.1073/pnas.0603565104.

[92] Flack JC, Erwin D, Elliot T, Krakauer DC. Timescales, symmetry, and uncertainty reduction in the origins of hierarchy in biological systems. In: Sterelny K, Joyce R, Calcott B, Fraser B, editors. Evolution and its cooperation. MIT Press; 2012.

[93] Fotopoulou A, Tsakiris M. Mentalizing homeostasis: the social origins of interoceptive inference. Neuropsychoanalysis 2017;19(1):3–28. https://doi.org/10.1080/15294145.2017.1294031.

[94] Foucault M. Discipline and punish: the birth of the prison. Knopf Doubleday Publishing Group; 2012. (Original work published 1975).

[95] Francis AL, Oliver J. Psychophysiological measurement of affective responses during speech perception. Hear Res 2018;369:103–19. https://doi.org/10.1016/j.heares.2018.07.007.

[96] Frank RH. Passions within reason: the strategic role of the emotions. Norton; 1988.

[97] Franklin DW, Wolpert DM. Computational mechanisms of sensorimotor control. Neuron 2011;72(3):425–42. https://doi.org/10.1016/j.neuron.2011.10.006.

[98] Fridman J, Barrett LF, Wormwood JB, Quigley KS. Applying the theory of constructed emotion to police decision making. Front Psychol 2019;10. https://doi.org/10.3389/fpsyg.2019.01946.

[99] Friston K. Hierarchical models in the brain. PLoS Comput Biol 2008;4(11):e1000211. https://doi.org/10.1371/journal.pcbi.1000211.

[100] Friston K. The free-energy principle: a unified brain theory?. Nat Rev Neurosci 2010;11(2):127–38. https://doi.org/10.1038/nrn2787.

[101] Friston K, FitzGerald T, Rigoli F, Schwartenbeck P, O'Doherty J, Pezzulo G. Active inference and learning. Neurosci Biobehav Rev 2016;68:862–79. https://doi.org/10.1016/j.neubiorev.2016.06.022.

[102] Friston K, FitzGerald T, Rigoli F, Schwartenbeck P, Pezzulo G. Active inference: a process theory. Neural Comput 2017;29(1):1–49. https://doi.org/10.1162/NECO_a_00912.

[103] Friston K, Rigoli F, Ognibene D, Mathys C, Fitzgerald T, Pezzulo G. Active inference and epistemic value. Cogn Neurosci 2015;6(4):187–214. https://doi.org/10.1080/17588928.2015.1020053.

[104] Friston K, Thornton C, Clark A. Free-energy minimization and the dark-room problem. Front Psychol 2012;3. https://doi.org/10.3389/fpsyg.2012.00130.

[105] Gailliot MT, Baumeister RF. The physiology of willpower: linking blood glucose to self-control. Personal Soc Psychol Rev 2007;11(4):303–27. https://doi.org/10.1177/1088868307303030.

[106] Gailliot MT, Baumeister RF, DeWall CN, Maner JK, Plant EA, Tice DM, et al. Self-control relies on glucose as a limited energy source: willpower is more than a metaphor. J Pers Soc Psychol 2007;92(2):325–36. https://doi.org/10.1037/0022-3514.92.2.325.

[107] Gallagher S. Understanding interpersonal problems in autism: interaction theory as an alternative to theory of mind. Philos Psychiatr Psychol 2004;11(3):199–217. https://doi.org/10.1353/ppp.2004.0063.

[108] Gallagher S. How the body shapes the mind. Clarendon Press; 2005.

[109] Gallagher S. Direct perception in the intersubjective context. Conscious Cogn 2008;17(2):535–43. https://doi.org/10.1016/j.concog.2008.03.003.

[110] Gallagher S. Decentering the brain: embodied cognition and the critique of neurocentrism and narrow-minded philosophy of mind. Constr Found 2018;14(1):8–21.

[111] Gallotti M, Fairhurst MT, Frith CD. Alignment in social interactions. Conscious Cogn 2017;48:253–61. https://doi.org/10.1016/j.concog.2016.12.002.

[112] Garrett JR. The proper role of nerves in salivary secretion: a review. J Dent Res 1987;66:387–97. https://doi.org/10.1177/00220345870660020201.

[113] Gasbarri A, Pompili A. Involvement of glutamate in learning and memory. In: Identification of neural markers accompanying memory. Elsevier; 2014. p. 63–77.

[114] Geanakoplos J, Pearce D, Stacchetti E. Psychological games and sequential rationality. Games Econ Behav 1989;1(1):60–79. https://doi.org/10.1016/0899-8256(89)90005-5.

[115] Gendron M, Roberson D, van der Vyver JM, Barrett LF. Cultural relativity in perceiving emotion from vocalizations. Psychol Sci 2014;25(4):911–20. https://doi.org/10.1177/0956797613517239.

[116] Gerrans P, Stone VE. Generous or parsimonious cognitive architecture? Cognitive neuroscience and theory of mind. Br J Philos Sci 2008;59(2):121–41. https://doi.org/10.1093/bjps/axm038.

[117] Gilbert DT. Ordinary personology. In: The handbook of social psychology. McGraw-Hill; 1998. p. 89–150.

[118] Gintis H. Gene–culture coevolution and the nature of human sociality. Philos Trans R Soc Lond B, Biol Sci 2011;366(1566):878–88. https://doi.org/10.1098/rstb.2010.0310.

[119] Godfrey-Smith P. Complexity and the function of mind in nature. Cambridge University Press; 1998.
[120] Godfrey-Smith P. Environmental complexity and the evolution of cognition. In: Sternberg R, Kaufman J, editors. The evolution of intelligence. Lawrence Erlbaum; 2002. p. 233–49.
[121] Godfrey-Smith P. Complexity revisited. Biol Philos 2017;32(3):467–79. https://doi.org/10.1007/s10539-017-9569-z.
[122] Goldman AI. Mirroring, simulating and mindreading. Mind Lang 2009;24(2):235–52. https://doi.org/10.1111/j.1468-0017.2008.01361.x.
[123] Goldman AI, Jordan L. Mindreading by simulation: the roles of imagination and mirroring. In: Baron-Cohen S, Lombardo M, Tager-Flusberg H, editors. Understanding other minds. 3rd ed. Oxford University Press; 2013. p. 448–66.
[124] Gomez P, von Gunten A, Danuser B. Autonomic nervous system reactivity within the valence-arousal affective space: modulation by sex and age. Int J Psychophysiol: Off J Int Organ Psychophysiol 2016;109:51–62. https://doi.org/10.1016/j.ijpsycho.2016.10.002.
[125] Gopnik A. The theory theory as an alternative to the innateness hypothesis. In: Antony LM, Hornstein N, editors. Chomsky and his critics. Blackwell Publishing Ltd.; 2003. p. 238–54.
[126] Gopnik A, Griffiths TL, Lucas CG. When younger learners can be better (or at least more open-minded) than older ones. Curr Dir Psychol Sci 2015;24(2):87–92. https://doi.org/10.1177/0963721414556653.
[127] Gopnik A, O'Grady S, Lucas CG, Griffiths TL, Wente A, Bridgers S, et al. Changes in cognitive flexibility and hypothesis search across human life history from childhood to adolescence to adulthood. Proc Natl Acad Sci 2017;114(30):7892–9. https://doi.org/10.1073/pnas.1700811114.
[128] Gopnik A, Wellman HM. Why the child's theory of mind really is a theory. Mind Lang 1992;7(1–2):145–71. https://doi.org/10.1111/j.1468-0017.1992.tb00202.x.
[129] Gopnik A, Wellman HM. Reconstructing constructivism: causal models, Bayesian learning mechanisms and the theory theory. Psychol Bull 2012;138(6):1085–108. https://doi.org/10.1037/a0028044.
[130] Gordon RM. The simulation theory: objections and misconceptions. Mind Lang 1992;7(1–2):11–34.
[131] Goyal MS, Hawrylycz M, Miller JA, Snyder AZ, Raichle ME. Aerobic glycolysis in the human brain is associated with development and neotenous gene expression. Cell Metab 2014;19(1):49–57. https://doi.org/10.1016/j.cmet.2013.11.020.
[132] Greenwald AG. The totalitarian ego: fabrication and revision of personal history. Am Psychol 1980;35(7):603–18. https://doi.org/10.1037/0003-066X.35.7.603.
[133] Greenwood JD. The disappearance of the social in American social psychology. Cambridge University Press; 2004.
[134] Grice HP. Studies in the way of words. Harvard University Press; 1991.
[135] Griffiths TL. Bayesian models as tools for exploring inductive biases. In: Banich MT, Caccamise D, editors. Generalization of knowledge: multidisciplinary perspectives. Psychology Press; 2010.
[136] Griffiths TL, Lieder F, Goodman ND. Rational use of cognitive resources: levels of analysis between the computational and the algorithmic. Top Cogn Sci 2015;7(2):217–29. https://doi.org/10.1111/tops.12142.
[137] Guillory SA, Bujarski KA. Exploring emotions using invasive methods: review of 60 years of human intracranial electrophysiology. Soc Cogn Affect Neurosci 2014;9(12):1880–9. https://doi.org/10.1093/scan/nsu002.
[138] Haidt J. The emotional dog and its rational tail: a social intuitionist approach to moral judgment. Psychol Rev 2001;108(4):814–34. https://doi.org/10.1037/0033-295X.108.4.814.
[139] Haidt J, Joseph C. Intuitive ethics: how innately prepared intuitions generate culturally variable virtues. Daedalus 2004;133(4):55–66.
[140] Hajcak G, McDonald N, Simons RF. To err is autonomic: error-related brain potentials, ANS activity, and post-error compensatory behavior. Psychophysiology 2003;40(6):895–903. https://doi.org/10.1111/1469-8986.00107.
[141] Hamilton WD. The genetical evolution of social behaviour. I. J Theor Biol 1964;7(1):1–16. https://doi.org/10.1016/0022-5193(64)90038-4.
[142] Harris JJ, Attwell D. The energetics of CNS white matter. J Neurosci 2012;32(1):356–71. https://doi.org/10.1523/JNEUROSCI.3430-11.2012.
[143] Hawkins RXD, Goodman ND, Goldstone RL. The emergence of social norms and conventions. Trends Cogn Sci 2019;23(2):158–69. https://doi.org/10.1016/j.tics.2018.11.003.
[144] Heider F. The psychology of interpersonal relations. John Wiley and Sons Inc.; 1958.
[145] Henrich J. The secret of our success: how culture is driving human evolution, domesticating our species, and making us smarter. Princeton University Press; 2015.
[146] Henrich J, Heine SJ, Norenzayan A. The weirdest people in the world?. Behav Brain Sci 2010;33(2–3):61–83. https://doi.org/10.1017/S0140525X0999152X.
[147] Henrich J, McElreath R. The evolution of cultural evolution. Evol Anthropol: Issues News Rev 2003;12(3):123–35. https://doi.org/10.1002/evan.10110.
[148] Hertz L, Gibbs ME. What learning in day-old chickens can teach a neurochemist: focus on astrocyte metabolism. J Neurochem 2009;109(Suppl. 1):10–6. https://doi.org/10.1111/j.1471-4159.2009.05939.x.
[149] Heyes C. Grist and mills: on the cultural origins of cultural learning. Philos Trans R Soc Lond B, Biol Sci 2012;367(1599):2181–91. https://doi.org/10.1098/rstb.2012.0120.
[150] Heyes C. Cognitive gadgets: the cultural evolution of thinking. Harvard University Press; 2018.
[151] Hoffman M, Yoeli E, Nowak MA. Cooperate without looking: why we care what people think and not just what they do. Proc Natl Acad Sci 2015;112(6):1727–32. https://doi.org/10.1073/pnas.1417904112.
[152] Hofman MA. Energy metabolism, brain size and longevity in mammals. Q Rev Biol 1983;58(4):495–512.
[153] Hoffman ML. Developmental synthesis of affect and cognition and its implications for altruistic motivation. Dev Psychol 1975;11(5):607–22. https://doi.org/10.1037/0012-1649.11.5.607.
[154] Hogg MA. Subjective uncertainty reduction through self-categorization: a motivational theory of social identity processes. Eur Rev Soc Psychol 2000;11(1):223–55. https://doi.org/10.1080/14792772043000040.

[155] Hogg MA. Uncertainty–identity theory. In: Advances in experimental social psychology, vol. 39. Academic Press; 2007. p. 69–126.

[156] Hohwy J. The predictive mind. Oxford University Press; 2013.

[157] Holroyd CB. The waste disposal problem of effortful control. In: Braver TS, editor. Motivation and cognitive control. Psychology Press; 2016. p. 235–60.

[158] Hutchinson B, Barrett LF. The power of predictions: an emerging paradigm for psychological research. Curr Dir Psychol Sci 2019;28(3):280–91. https://doi.org/10.1177/0963721419831992.

[159] James W. The principles of psychology (vol. 1). Holt; 1931. Available from: http://archive.org/details/theprinciplesofp01jameuoft. (Original work published 1890).

[160] Jebari J. Empirical moral rationalism and the social constitution of normativity. Philos Stud 2020;1(25). https://doi.org/10.1007/s11098-018-1134-3 [in press].

[161] Jones EE, Davis KE. From acts to dispositions: the attribution process in person perception. In: Berkowitz L, editor. Advances in experimental social psychology, vol. 2. Academic Press; 1965. p. 219–66.

[162] Jordan JJ, Hoffman M, Bloom P, Rand DG. Third-party punishment as a costly signal of trustworthiness. Nature 2016;530(7591):473–6. https://doi.org/10.1038/nature16981.

[163] Kahneman D, Knetsch JL, Thaler RH. Anomalies: the endowment effect, loss aversion, and status quo bias. J Econ Perspect 1991;5(1):193–206. https://doi.org/10.1257/jep.5.1.193.

[164] Kant I. Groundwork of the metaphysics of morals. Cambridge University Press; 1998. (Original work published 1785).

[165] Kant I. The critique of pure reason. Project Gutenberg Literary Archive Foundation 2003. [Aldarondo C, Widger D, editors, Meiklejohn JMD, trans.]. Available from: https://www.gutenberg.org/files/4280/4280-h/4280-h.htm. (Original work published 1781).

[166] Keil A, Freund AM. Changes in the sensitivity to appetitive and aversive arousal across adulthood. Psychol Aging 2009;24(3):668–80. https://doi.org/10.1037/a0016969.

[167] Kelley HH. Two functions of reference groups. In: Swanson GE, Newcomb TM, Hartley EL, editors. Readings in social psychology. 2nd ed. Holt, Rinehart and Winston; 1952. p. 410–4.

[168] Kelley HH. Attribution theory in social psychology. In: Levine D, editor. Nebraska symposium on motivation, vol. 15. University of Nebraska Press; 1967. p. 192–238.

[169] Kennedy C, Sokoloff L. An adaptation of the nitrous oxide method to the study of the cerebral circulation in children; normal values for cerebral blood flow and cerebral metabolic rate in childhood. J Clin Invest 1957;36(7):1130–7. https://doi.org/10.1172/JCI103509.

[170] Khalsa SS, Adolphs R, Cameron OG, Critchley HD, Davenport PW, Feinstein JS, et al. Interoception and mental health: a roadmap. Biol Psychiatry Cogn Neurosci Neuroimaging 2018;3(6):501–13. https://doi.org/10.1016/j.bpsc.2017.12.004.

[171] King Jr ML. Letter from a Birmingham jail; 1963. Available from: https://www.africa.upenn.edu/Articles_Gen/Letter_Birmingham.html.

[172] Kinzler KD, Shutts K, DeJesus J, Spelke ES. Accent trumps race in guiding children's social preferences. Social Cogn 2009;27(4):623–34. https://doi.org/10.1521/soco.2009.27.4.623.

[173] Kirby KN, Maraković NN. Delay-discounting probabilistic rewards: rates decrease as amounts increase. Psychon Bull Rev 1996;3(1):100–4. https://doi.org/10.3758/BF03210748.

[174] Kleckner IR, Zhang J, Touroutoglou A, Chanes L, Xia C, Simmons WK, et al. Evidence for a large-scale brain system supporting allostasis and interoception in humans. Nat Hum Behav 2017;1(5):0069. https://doi.org/10.1038/s41562-017-0069.

[175] Kleiber M. Body size and metabolism. Hilgardia 1932;6(11):315–53.

[176] Kohlberg L. From is to ought: how to commit the naturalistic fallacy and get away with it in the study of moral development. In: Mischel T, editor. Cognitive development and epistemology. Academic Press; 1971. p. 151–235.

[177] Koster-Hale J, Saxe R. Theory of mind: a neural prediction problem. Neuron 2013;79(5):836–48. https://doi.org/10.1016/j.neuron.2013.08.020.

[178] Kozak MN, Marsh AA, Wegner DM. What do I think you're doing? Action identification and mind attribution. J Pers Soc Psychol 2006;90(4):543–55. https://doi.org/10.1037/0022-3514.90.4.543.

[179] Kriss PH, Weber RA, Xiao E. Turning a blind eye, but not the other cheek: on the robustness of costly punishment. J Econ Behav Organ 2016;128:159–77. https://doi.org/10.1016/j.jebo.2016.05.017.

[180] Kuhn TS. The structure of scientific revolutions. University of Chicago Press; 2012. (Original work published 1962).

[181] Kunda Z. The case for motivated reasoning. Psychol Bull 1990;108(3):480–98. https://doi.org/10.1037/0033-2909.108.3.480.

[182] Kuppens P, Tuerlinckx F, Russell JA, Barrett LF. The relation between valence and arousal in subjective experience. Psychol Bull 2013;139(4):917–40. https://doi.org/10.1037/a0030811.

[183] Kurzban R, Burton-Chellew MN, West SA. The evolution of altruism in humans. Annu Rev Psychol 2015;66(1):575–99. https://doi.org/10.1146/annurev-psych-010814-015355.

[184] Kurzban R, Duckworth A, Kable JW, Myers J. An opportunity cost model of subjective effort and task performance. Behav Brain Sci 2013;36(6). https://doi.org/10.1017/S0140525X12003196.

[185] Lange F, Eggert F. Sweet delusion. Glucose drinks fail to counteract ego depletion. Appetite 2014;75:54–63. https://doi.org/10.1016/j.appet.2013.12.020.

[186] Leslie AM. Pretense and representation: the origins of "theory of mind". Psychol Rev 1987;94(4):412–26.

[187] Leslie AM, German TP, Polizzi P. Belief-desire reasoning as a process of selection. Cogn Psychol 2005;50(1):45–85. https://doi.org/10.1016/j.cogpsych.2004.06.002.

[188] Lieder F, Griffiths TL. Resource-rational analysis: understanding human cognition as the optimal use of limited computational resources. Behav Brain Sci 2019:1–85. https://doi.org/10.1017/S0140525X1900061X.

[189] Lindquist KA, Gendron M. What's in a word? Language constructs emotion perception. Emot Rev 2013;5(1):66–71. https://doi.org/10.1177/1754073912451351.

[190] Lucas CG, Bridgers S, Griffiths TL, Gopnik A. When children are better (or at least more open-minded) learners than adults: developmental differences in learning the forms of causal relationships. Cognition 2014;131(2):284–99. https://doi.org/10.1016/j.cognition.2013.12.010.

[191] Malle BF. Attribution theories: how people make sense of behavior. In: Chadee D, editor. Theories in social psychology. Wiley-Blackwell; 2011. p. 72–96.

[192] Mameli M. Mindreading, mindshaping, and evolution. Biol Philos 2001;16(5):595–626. https://doi.org/10.1023/A:1012203830990.

[193] Mason JW. A re-evaluation of the concept of 'non-specificity' in stress theory. J Psychiatr Res 1971;8(3):323–33. https://doi.org/10.1016/0022-3956(71)90028-8.

[194] Mather M, Clewett D, Sakaki M, Harley CW. Norepinephrine ignites local hotspots of neuronal excitation: how arousal amplifies selectivity in perception and memory. Behav Brain Sci 2016;39:e200. https://doi.org/10.1017/S0140525X15000667.

[195] McCormick CM, Mathews IZ, Thomas C, Waters P. Investigations of HPA function and the enduring consequences of stressors in adolescence in animal models. Brain Cogn 2010;72(1):73–85. https://doi.org/10.1016/j.bandc.2009.06.003.

[196] McGeer V. The regulative dimension of folk-psychology. In: Hutto D, Ratcliffe M, editors. Folk-psychology reassessed. Springer; 2007.

[197] McGeer V. Mind-making practices: the social infrastructure of self-knowing agency and responsibility. Philos Explor 2015;18(2):259–81. https://doi.org/10.1080/13869795.2015.1032331.

[198] McNamara RA, Willard AK, Norenzayan A, Henrich J. Weighing outcome vs. intent across societies: how cultural models of mind shape moral reasoning. Cognition 2018;182:95–108. https://doi.org/10.1016/j.cognition.2018.09.008.

[199] Mergenthaler P, Lindauer U, Dienel GA, Meisel A. Sugar for the brain: the role of glucose in physiological and pathological brain function. Trends Neurosci 2013;36(10):587–97. https://doi.org/10.1016/j.tins.2013.07.001.

[200] Mesulam M. From sensation to cognition. Brain 1998;121(6):1013–52. https://doi.org/10.1093/brain/121.6.1013.

[201] Milgram S. Behavioral study of obedience. J Abnorm Psychol 1963;67:371–8.

[202] Moreno A, Lasa A. From basic adaptivity to early mind: the origin and evolution of cognitive capacities. Evol Cogn 2003;9(1):12–30.

[203] Moreno A, Mossio M. Biological autonomy, vol. 12. Springer Netherlands; 2015.

[204] Morris A, Cushman F. A common framework for theories of norm compliance. Soc Philos Policy 2018;35(1):101–27. https://doi.org/10.1017/S0265052518000134.

[205] Moscovici S. Social influence and social change. Academic Press; 1976.

[206] Moscovici S. Toward a theory of conversion behavior. In: Berkowitz L, editor. Advances in experimental social psychology, vol. 13. Academic Press; 1980. p. 209–39.

[207] Moscovici S, Zavalloni M. The group as a polarizer of attitudes. J Pers Soc Psychol 1969;12(2):125–35. https://doi.org/10.1037/h0027568.

[208] Nettle D. The cultural and the agentic. In: Hanging on to the edges: essays on science, society, and the academic life. Open Book Publishers; 2018. p. 43–58.

[209] Niven JE. Neuronal energy consumption: biophysics, efficiency and evolution. Curr Opin Neurobiol 2016;41:129–35. https://doi.org/10.1016/j.conb.2016.09.004.

[210] Niven JE, Laughlin SB. Energy limitation as a selective pressure on the evolution of sensory systems. J Exp Biol 2008;211(11):1792–804. https://doi.org/10.1242/jeb.017574.

[211] Nowak MA. Five rules for the evolution of cooperation. Science 2006;314(5805):1560–3. https://doi.org/10.1126/science.1133755.

[212] Olin L. Questions for a theory of humor. Philos Compass 2016;11(6):338–50. https://doi.org/10.1111/phc3.12320.

[213] Ondobaka S, Kilner J, Friston K. The role of interoceptive inference in theory of mind. Brain Cogn 2017;112:64–8. https://doi.org/10.1016/j.bandc.2015.08.002.

[214] Open Science Collaboration. Estimating the reproducibility of psychological science. Science 2015;349(6251):aac4716. https://doi.org/10.1126/science.aac4716.

[215] Orquin JL, Kurzban R. A meta-analysis of blood glucose effects on human decision making. Psychol Bull 2016;142(5):546–67. https://doi.org/10.1037/bul0000035.

[216] Paluck EL. How to overcome prejudice. Science 2016;352(6282):147. https://doi.org/10.1126/science.aaf5207.

[217] Paluck EL, Shepherd H, Aronow PM. Changing climates of conflict: a social network experiment in 56 schools. Proc Natl Acad Sci 2016;113(3):566–71. https://doi.org/10.1073/pnas.1514483113.

[218] Pavlov I. Nobel lecture: physiology of digestion. NobelPrize.Org 2018. Available from: https://www.nobelprize.org/prizes/medicine/1904/pavlov/lecture/. (Original work published 1904).

[219] Pedersen EJ, Kurzban R, McCullough ME. Do humans really punish altruistically? A closer look. Proc R Soc Lond B, Biol Sci 2013;280(1758). https://doi.org/10.1098/rspb.2012.2723.

[220] Pedersen EJ, McAuliffe WHB, McCullough ME. The unresponsive avenger: more evidence that disinterested third parties do not punish altruistically. J Exp Psychol Gen 2018;147(4):514–44. https://doi.org/10.1037/xge0000410.

[221] Piliavin JA, Dovidio JF, Gaertner SL, Clark RD. Emergency intervention. Academic Press; 1981.

[222] Pontzer H. Energy expenditure in humans and other primates: a new synthesis. Annu Rev Anthropol 2015;44(1):169–87. https://doi.org/10.1146/annurev-anthro-102214-013925.

[223] Premack D, Woodruff G. Does the chimpanzee have a theory of mind?. Behav Brain Sci 1978;1(4):515–26. https://doi.org/10.1017/S0140525X00076512.

[224] Preuschoff K, 't Hart BM, Einhäuser W. Pupil dilation signals surprise: evidence for noradrenaline's role in decision making. Front Neurosci 2011;5. https://doi.org/10.3389/fnins.2011.00115.

[225] Quine WVO, Ullian JS. The web of belief. Random House; 1970.

[226] Raichle ME. The restless brain: how intrinsic activity organizes brain function. Philos Trans R Soc Lond B, Biol Sci 2015;370(1668):20140172. https://doi.org/10.1098/rstb.2014.0172.

[227] Raichle ME, Gusnard DA. Appraising the brain's energy budget. Proc Natl Acad Sci USA 2002;99(16):10237–9. https://doi.org/10.1073/pnas.172399499.

[228] Railton P. The affective dog and its rational tale: intuition and attunement. Ethics 2014;124(4):813–59. https://doi.org/10.1086/675876.

[229] Rand DG, Yoeli E, Hoffman M. Harnessing reciprocity to promote cooperation and the provisioning of public goods. Policy Insights Behav Brain Sci 2014;1(1):263–9. https://doi.org/10.1177/2372732214548426.

[230] Rao RPN, Ballard DH. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nat Neurosci 1999;2(1):79–87. https://doi.org/10.1038/4580.

[231] Reber R, Norenzayan A. Shared fluency theory of social cohesiveness: how the metacognitive feeling of processing fluency contributes to group processes. In: Metacognitive diversity: an interdisciplinary approach. Oxford University Press; 2018. p. 47–67.

[232] Reber R, Schwarz N, Winkielman P. Processing fluency and aesthetic pleasure: is beauty in the perceiver's processing experience?. Personal Soc Psychol Rev 2004;8(4):364–82. https://doi.org/10.1207/s15327957pspr0804_3.

[233] Richardson H, Saxe R. Development of predictive responses in theory of mind brain regions. Dev Sci 2019;e12863. https://doi.org/10.1111/desc.12863.

[234] Richerson PJ, Boyd R. Not by genes alone: how culture transformed human evolution. University of Chicago Press; 2008.

[235] Ross Ashby W. Design for a brain: the origin of adaptive behavior. Chapman and Hall; 1960.

[236] Ross Ashby W. The brain as regulator. Nature 1960;186:413. https://doi.org/10.1038/186413a0.

[237] Ross D. Game theory. In: Zalta EN, editor. The Stanford encyclopedia of philosophy; 2018. Available from: https://plato.stanford.edu/archives/fall2018/entries/game-theory/.

[238] Russek EM, Momennejad I, Botvinick MM, Gershman SJ, Daw ND. Predictive representations can link model-based reinforcement learning to model-free mechanisms. PLoS Comput Biol 2017;13(9):e1005768. https://doi.org/10.1371/journal.pcbi.1005768.

[239] Russell JA. A circumplex model of affect. J Pers Soc Psychol 1980;39(6):1161–78. https://doi.org/10.1037/h0077714.

[240] Russell JA. Core affect and the psychological construction of emotion. Psychol Rev 2003;110(1):145–72. https://doi.org/10.1037/0033-295X.110.1.145.

[241] Russell JA, Barrett LF. Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant. J Pers Soc Psychol 1999;76(5):805–19. https://doi.org/10.1037/0022-3514.76.5.805.

[242] Sadock JM, Zwicky AM. Speech acts distinctions in syntax. In: Shopen T, editor. Language typology and syntactic description. Cambridge University Press; 1985. p. 155–96.

[243] Samuelson PA. A note on the pure theory of consumer's behaviour. Economica 1938;5(17):61–71. https://doi.org/10.2307/2548836.

[244] Samuelson W, Zeckhauser R. Status quo bias in decision making. J Risk Uncertain 1988;1(1):7–59. https://doi.org/10.1007/BF00055564.

[245] Sands M, Garbacz A, Isaacowitz DM. Just change the channel? Studying effects of age on emotion regulation using a TV watching paradigm. Soc Psychol Pers Sci 2016;7(8):788–95. https://doi.org/10.1177/1948550616660593.

[246] Sands M, Isaacowitz DM. Situation selection across adulthood: the role of arousal. Cogn Emot 2017;31(4):791–8. https://doi.org/10.1080/02699931.2016.1152954.

[247] Schaafsma SM, Pfaff DW, Spunt RP, Adolphs R. Deconstructing and reconstructing theory of mind. Trends Cogn Sci 2015;19(2):65–72. https://doi.org/10.1016/j.tics.2014.11.007.

[248] Scholl BJ, Leslie AM. Minds, modules, and meta-analysis. Child Dev 2001;72(3):696–701. https://doi.org/10.1111/1467-8624.00308.

[249] Schulkin J. Social allostasis: anticipatory regulation of the internal milieu. Front Evol Neurosci 2011;2. https://doi.org/10.3389/fnevo.2010.00111.

[250] Schulz JF, Bahrami-Rad D, Beauchamp JP, Henrich J. The Church, intensive kinship, and global psychological variation. Science 2019;366(6466). https://doi.org/10.1126/science.aau5141.

[251] Schwartz SH. Normative influences on altruism. In: Berkowitz L, editor. Advances in experimental social psychology, vol. 10. Academic Press; 1977. p. 221–79.

[252] Schwartz SH, Gottlieb A. Bystander reactions to a violent theft: crime in Jerusalem. J Pers Soc Psychol 1976;34(6):1188–99.

[253] Schwartz SH, Gottlieb A. Bystander anonymity and reactions to emergencies. J Pers Soc Psychol 1980;39(3):418–30. https://doi.org/10.1037/0022-3514.39.3.418.

[254] Searle JR. The rediscovery of the mind. MIT Press; 1992.

[255] Searle JR. The construction of social reality. The Free Press; 1995.

[256] Searle JR. Mind: a brief introduction. Oxford University Press; 2004.

[257] Searle JR. Making the social world: the structure of human civilization. Oxford University Press; 2010.

[258] Seiver E, Gopnik A, Goodman ND. Did she jump because she was the big sister or because the trampoline was safe? Causal inference and the development of social attribution. Child Dev 2013;84(2):443–54. https://doi.org/10.1111/j.1467-8624.2012.01865.x.

[259] Sengupta B, Stemmler MB, Friston KJ. Information and efficiency in the nervous system—a synthesis. PLoS Comput Biol 2013;9(7):e1003157. https://doi.org/10.1371/journal.pcbi.1003157.

[260] Sengupta B, Stemmler M, Laughlin SB, Niven JE. Action potential energy efficiency varies among neuron types in vertebrates and invertebrates. PLoS Comput Biol 2010;6(7):e1000840. https://doi.org/10.1371/journal.pcbi.1000840.

[261] Seth AK. Interoceptive inference, emotion, and the embodied self. Trends Cogn Sci 2013;17(11):565–73. https://doi.org/10.1016/j.tics.2013.09.007.

[262] Seth AK. The cybernetic bayesian brain: from interoceptive inference to sensorimotor contingencies. In: Metzinger T, Windt JM, editors. Open MIND. MIND Group; 2015. p. 1–24. Available from: http://www.open-mind.net/DOI?isbn=9783958570108.

[263] Seth AK, Friston KJ. Active interoceptive inference and the emotional brain. Philos Trans R Soc Lond B, Biol Sci 2016;371(1708). https://doi.org/10.1098/rstb.2016.0007.

[264] Shadmehr R, Krakauer JW. A computational neuroanatomy for motor control. Exp Brain Res 2008;185(3):359–81. https://doi.org/10.1007/s00221-008-1280-5.

[265] Shadmehr R, Smith MA, Krakauer JW. Error correction, sensory prediction, and adaptation in motor control. Annu Rev Neurosci 2010;33(1):89–108. https://doi.org/10.1146/annurev-neuro-060909-153135.

[266] Shannon C, Weaver W. The mathematical theory of communication. 10th ed. The University of Illinois Press; 1964. (Original work published 1949).

[267] Sheahan HR, Franklin DW, Wolpert DM. Motor planning, not execution, separates motor memories. Neuron 2016;92(4):773–9. https://doi.org/10.1016/j.neuron.2016.10.017.

[268] Smith A. The theory of moral sentiments. In: Hanley RP, editor. Penguin Classics; 2010. (Original work published 1790).

[269] Sokoloff L, Mangold R, Wechsler RL, Kennedy C, Kety SS. The effect of mental arithmetic on cerebral circulation and metabolism. J Clin Invest 1955;34(7 Pt 1):1101–8.

[270] Sokoloff L, Reivich M, Kennedy C, Rosiers MHD, Patlak CS, Pettigrew KD, et al. The [14c] deoxyglucose method for the measurement of local cerebral glucose utilization: theory, procedure, and normal values in the conscious and anesthetized albino rat. J Neurochem 1977;28(5):897–916. https://doi.org/10.1111/j.1471-4159.1977.tb10649.x.

[271] Sommer MA, Wurtz RH. What the brain stem tells the frontal cortex. I. Oculomotor signals sent from superior colliculus to frontal eye field via mediodorsal thalamus. J Neurophysiol 2004;91(3):1381–402. https://doi.org/10.1152/jn.00738.2003.

[272] Sommer MA, Wurtz RH. What the brain stem tells the frontal cortex. II. Role of the SC-MD-FEF pathway in corollary discharge. J Neurophysiol 2004;91(3):1403–23. https://doi.org/10.1152/jn.00740.2003.

[273] Sperry RW. Neural basis of the spontaneous optokinetic response produced by visual inversion. J Comp Physiol Psychol 1950;43(6):482–9.

[274] Spivey M. The continuity of mind. Oxford University Press; 2008.

[275] Spratling MW. A review of predictive coding algorithms. Brain Cogn 2017;112:92–7. https://doi.org/10.1016/j.bandc.2015.11.003.

[276] Spruit IM, Wilderjans TF, van Steenbergen H. Heart work after errors: behavioral adjustment following error commission involves cardiac effort. Cogn Affect Behav Neurosci 2018;18(2):375–88. https://doi.org/10.3758/s13415-018-0576-6.

[277] Stel M, Blascovich J, McCall C, Mastop J, van Baaren RB, Vonk R. Mimicking disliked others: effects of a priori liking on the mimicry-liking link. Eur J Soc Psychol 2009. https://doi.org/10.1002/ejsp.655.

[278] Sterling P. Allostasis: a model of predictive regulation. Physiol Behav 2012;106(1):5–15. https://doi.org/10.1016/j.physbeh.2011.06.004.

[279] Sterling P. Predictive regulation and human design. eLife 2018;7. https://doi.org/10.7554/eLife.36133.

[280] Sterling P, Eyer J. Allostasis: a new paradigm to explain arousal pathology. In: Fisher S, Reason J, editors. Handbook of life stress, cognition and health. John Wiley and Sons; 1988. p. 629–49.

[281] Sterling P, Laughlin S. Principles of neural design. MIT Press; 2015.

[282] Theriault JE, Waytz A, Heiphetz L, Young LL. Theory of Mind network activity is associated with metaethical judgment: an item analysis. PsyArXiv 2017. https://doi.org/10.31234/osf.io/gb5am [submitted for publication].

[283] Theriault JE, Young LL. Taking an "intentional stance" on moral psychology. In: Systma J, editor. Advances in experimental philosophy of mind. Bloomsbury Publishing; 2014. p. 101–24.

[284] Thompson-Schill SL, Ramscar M, Chrysikou EG. Cognition without control: when a little frontal lobe goes a long way. Curr Dir Psychol Sci 2009;18(5):259–63. https://doi.org/10.1111/j.1467-8721.2009.01648.x.

[285] Toelch U, Dolan RJ. Informational and normative influences in conformity from a neurocomputational perspective. Trends Cogn Sci 2015;19(10):579–89. https://doi.org/10.1016/j.tics.2015.07.007.

[286] Tomasello M. The moral psychology of obligation. Behav Brain Sci 2019:1–33. https://doi.org/10.1017/S0140525X19001742 [in press].

[287] Tooby J, Cosmides L. The psychological foundations of culture. In: Barkow J, Cosmides L, Tooby J, editors. The adapted mind: evolutionary psychology and the generation of culture. Oxford University Press; 1992. p. 19–136.

[288] Trivers RL. The evolution of reciprocal altruism. Q Rev Biol 1971;46(1):35–57. https://doi.org/10.1086/406755.

[289] Trivers RL. Foreword. In: Dawkins R, editor. The selfish gene: 40th anniversary edition. Oxford University Press; 2016. (Original work published 1976).

[290] Uttal WR. The new phrenology: the limits of localizing cognitive processes in the brain. MIT Press; 2001.

[291] Vallacher RR, Wegner DM. What do people think they're doing? Action identification and human behavior. Psychol Rev 1987;94(1):3–15. https://doi.org/10.1037/0033-295X.94.1.3.

[292] van den Berg R, Ma WJ. A resource-rational theory of set size effects in human visual working memory. eLife 2018;7. https://doi.org/10.7554/eLife.34963.

[293] von Helmholtz H. Treatise on physiological optics, vol. III. In: Southall JPC, editor, 3rd ed. Leopold Voss; 1910. Available from: http://echo.mpiwg-berlin.mpg.de/ECHOdocuView?url=/permanent/library/HS7FH69N/pageimg&viewMode=image&mode=imagepath&pn=7. (Original work published 1867).

[294] von Hippel W, Trivers R. The evolution and psychology of self-deception. Behav Brain Sci 2011;34(1):1–16. https://doi.org/10.1017/S0140525X10001354.

[295] von Hippel, von Holst E. Relations between the central nervous system and the peripheral organs. Br J Anim Behav 1954;2(3):89–94. https://doi.org/10.1016/S0950-5601(54)80044-X.

[296] Warnell KR, Redcay E. Minimal coherence among varied theory of mind measures in childhood and adulthood. Cognition 2019;191:103997. https://doi.org/10.1016/j.cognition.2019.06.009.

[297] Waytz A, Morewedge CK, Epley N, Monteleone G, Gao J-H, Cacioppo JT. Making sense by making sentient: effectance motivation increases anthropomorphism. J Pers Soc Psychol 2010;99(3):410–35. https://doi.org/10.1037/a0020240.

[298] Weibel ER. Symmorphosis: on form and function in shaping life. Harvard University Press; 2000.

[299] Weiss JM. Effects of coping behavior in different warning signal conditions on stress pathology in rats. J Comp Physiol Psychol 1971;77(1):1–13. https://doi.org/10.1037/h0031583.

[300] Westbrook A, Braver TS. Cognitive effort: a neuroeconomic approach. Cogn Affect Behav Neurosci 2015;15(2):395–415. https://doi.org/10.3758/s13415-015-0334-y.

[301] Westermann G, Mareschal D, Johnson MH, Sirois S, Spratling MW, Thomas MSC. Neuroconstructivism. Dev Sci 2007;10(1):75–83. https://doi.org/10.1111/j.1467-7687.2007.00567.x.

[302] Wilkins JS, Bourrat P. Replication and reproduction. In: Zalta EN, editor. The Stanford encyclopedia of philosophy (summer 2019). Meta-physics Research Lab, Stanford University; 2019. Available from: https://plato.stanford.edu/archives/sum2019/entries/replication/.

[303] Wiltermuth SS, Heath C. Synchrony and cooperation. Psychol Sci 2009;20(1):1–5. https://doi.org/10.1111/j.1467-9280.2008.02253.x.

[304] Wolpert DM, Flanagan JR. Computations underlying sensorimotor learning. Curr Opin Neurobiol 2016;37:7–11. https://doi.org/10.1016/j.conb.2015.12.003.

[305] Wood W, Lundgren S, Ouellette JA, Busceme S, Blackstone T. Minority influence: a meta-analytic review of social influence processes. Psychol Bull 1994;115(3):323–45.

[306] Wundt W. Outlines of psychology. Gustav E. Stechert; 1896 [Judd CH, trans.].

[307] Yang Y, Cao P, Yang Y, Wang S-R. Corollary discharge circuits for saccadic modulation of the pigeon visual system. Nat Neurosci 2008;11(5):595–602. https://doi.org/10.1038/nn.2107.

[308] Yoshida N. On reward function for survival. In: Joint 8th international conference on soft computing and intelligent systems and 17th international symposium on advanced intelligent systems; 2016. Available from: https://arxiv.org/abs/1606.05767v2.

[309] Yu AJ, Dayan P. Uncertainty, neuromodulation, and attention. Neuron 2005;46(4):681–92. https://doi.org/10.1016/j.neuron.2005.04.026.

[310] Zawidzki TW. The function of folk psychology: mind reading or mind shaping?. Philos Explor 2008;11(3):193–210. https://doi.org/10.1080/13869790802239235.

[311] Zawidzki TW. Mindshaping. In: Newen A, de Bruin L, Gallagher S, editors. Oxford handbook of 4e cognition. Oxford University Press; 2018.

[312] Zénon A, Solopchuk O, Pezzulo G. An information-theoretic perspective on the costs of cognition. Neuropsychologia 2019;123(4):5–18. https://doi.org/10.1016/j.neuropsychologia.2018.09.013.