Behavioral and Brain Sciences (forthcoming)

Thinking Through Other Minds: A Variational Approach to Cognition and Culture

Authors:

Samuel P. L. Veissière[1,2,3] (email: samuel.veissiere@mcgill.ca)

Axel Constant[3,4,5] (email: axel.constant.pruvost@gmail.com)

Maxwell J. D. Ramstead[1,3,5,6] (email: maxwell.ramstead@mail.mcgill.ca)

Karl J. Friston[5] (email: k.friston@ucl.ac.uk)

Laurence J. Kirmayer[1,2,3*] (laurence.kirmayer@mcgill.ca)

Affiliations:

1. Division of Social and Transcultural Psychiatry, Department of Psychiatry, McGill University, 1033 Pine Avenue, Montreal, QC, Canada.

2. Department of Anthropology, McGill University, 855 Sherbrooke Street West, H3A 2T7, Montreal, QC, Canada.

3. Culture, Mind, and Brain Program, McGill University, 1033 Pine Avenue, Montreal, QC, Canada.

4. Charles Perkins Centre, University of Sydney, John Hopkins, Camperdown NSW 2006, Australia.

5. Wellcome Trust Centre for Human Neuroimaging, University College London, London, WC1N 3BG, United Kingdom.

6. Department of Philosophy, McGill University, 855 Sherbrooke Street West, H3A 2T7, Montreal, QC, Canada.

*Corresponding author

Short abstract:

Notions such as ‘shared expectations’, the ‘selective patterning of attention and behaviour’, ‘cultural evolution’, ‘cultural inheritance’, and ‘implicit learning’ are the main candidates on which to base a unified account of social cognition and the acquisition of culture. However, they all require greater specification and clarification of how they interact. We integrate these candidates using the variational (free energy) approach to human cognition and culture in theoretical neuroscience. We show how human agents are able to learn shared expectations and norms through the selective patterning of attention and through the construction of social niches that afford epistemic resources (i.e., cultural affordances). We call this process “Thinking through Other Minds” (TTOM).

Long abstract:

The processes underwriting the acquisition of culture remain unclear. How are shared habits, norms, and expectations learned and maintained with precision and reliability across large-scale sociocultural ensembles? Is there a unifying account of the mechanisms involved in the acquisition of culture? Notions such as ‘shared expectations’, the ‘selective patterning of attention and behaviour’, ‘cultural evolution’, ‘cultural inheritance’, and ‘implicit learning’ are the main candidates to underpin a unifying account of cognition and the acquisition of culture; however, their interactions require greater specification and clarification. In this paper, we integrate these candidates using the variational (free energy) approach to human cognition and culture in theoretical neuroscience. We describe the construction by humans of social niches that afford epistemic resources called *cultural affordances*. We argue that human agents learn the shared habits, norms, and expectations of their culture through immersive participation in patterned cultural practices that selectively pattern attention and behaviour. We call this process “Thinking through Other Minds” (TTOM) – in effect, the process of

inferring other agents' expectations about the world and how to behave in social context. We argue that for humans, information *from and about other people's expectations* constitutes the primary domain of statistical regularities that humans leverage to predict and organize behaviour. The integrative model we offer has implications that can advance theories of cognition, enculturation, adaptation, and psychopathology. Crucially, this formal (variational) treatment seeks to resolve key debates in current cognitive science, such as the distinction between internalist and externalist accounts of Theory of Mind abilities and the more fundamental distinction between dynamical and representational accounts of enactivism.

Keywords: Cognition and culture; Variational free energy principle; Social learning; Epistemic Affordances; Cultural affordances; Niche construction; Embodiment; Enactment

> *[Humans] form with others joint goals to which both parties are normatively committed, they establish with others domains of joint attention and common conceptual ground, and they create with others symbolic, institutional realities that assign deontic powers to otherwise inert entities.* Michael Tomasello (Tomasello 2009), p. 105

> Choosing a swimsuit—
> when did his eyes replace mine?
> (mizugi erabu itsu shika kare no me to natte)
> Mayuzumi Madoka (Madoka 2003), p. xxxvi[1]

1. Introduction: Learning in cultural context

1.1. The puzzle of implicit cultural learning

Since the advent of the social sciences in the late 19th century, a recurring trope casts 'society' or, in its Durkheimian formulation, 'regulatory social forces' (Durkheim 1985/2014) as superordinate to individual human agency. As the story goes, humans acquire norms, tastes, preferences, and ways of doing things that are consistent with those of others in their local world and communities; that is, the relevant social and cultural groups (ingroups and outgroups) to which they belong and with whom they interact (Kurzban and Neuberg 2005).

Group variations in learned and structured dispositions extend to such domains as culturally shaped body practices like walking, sitting, eating, and sleeping (Mauss 1973), differentiated patterns of prejudice or bias against certain kinds of persons (e.g., racism, sexism, classism) (Machery 2016), proneness to optical illusions (McCauley and Henrich 2006), colour

perception (Goldstein, Davidoff, and Roberson 2009), food preferences (Wright, Nancarrow, and Kwok 2001), desirable body types (Swami et al. 2010), as well as thresholds for pain (Zatzick and Dimsdale 1990) and other forms of suffering and affliction that are shaped by culture (Kirmayer 1989; Kirmayer and Young 1998; Kirmayer, Gomez-Carrillo, and Veissière 2017), and historical context (Hacking 1998; Gold and Gold 2015). As developmental psychologists have argued, it is precisely because of the existence of inter-group behavioural and cognitive variations that arise through social learning within members of the same species that we can speak of culture (Tomasello 2009). We know there is such a 'thing' as culture, in other words, because there are cultural differences (Brown 2004). While it is clear that specific developmental experiences — governed by explicit social norms and contexts — shape these perceptual, cognitive, and attitudinal processes, most of cultural learning appears to be *implicit*, in the sense that it occurs without explicit instruction.

Implicit cultural learning poses a classical 'poverty of stimulus' problem, in that acquired knowledge, attitudes, and dispositions appear to go far beyond what can be learned by direct experience (Berwick and Chomsky 2013; Chomsky 1996) — they evince a special, ampliative form of abductive inference. For instance, alongside the many rules and facts about the world that are explicitly taught, human children learn a large and stable set of implicit beliefs that govern action without needing to be stated explicitly, described or explained (Sperber 1996, 1997). By age 7, children are already proficient in complex, though mostly tacit intergroup relational rules and dynamics of power, and already form implicit judgments about the 'value' of members of other groups, and that of their group in relation to others (e.g., children of minority groups often internalize preferences for prestige-laden groups different from their own ethnic group (for a review, see Machery and Faucher 2017; Kelly, Faucher, and Machery 2010; Pauker, Williams, and Steele 2016; Kinzler and Spelke 2011; Navarrete and

Fessler 2005; Clark and Clark 1939; Clark 1988; Edouard Machery and Faucher 2017; Huneman and Machery 2015))

Clearly, we are continuously immersed in culturally shaped environments and interactions from before birth. Despite advances in developmental psychology (Csibra and Gergely 2009; Tomasello 2014) and cognitive anthropology (Boyd and Richerson 2005), we still lack a formal account of the mechanisms of enculturation. The processes that enable implicit cultural habits and norms to arise from inference and imitation, and to be learned and maintained with a high degree of precision and reliability across large-scale sociocultural phenomena, involving multiple interlocking minds and institutional structures, are only partly understood. This is our puzzle.

1.2. The Theory of Mind debates

In this paper, we will propose a solution to the puzzle of implicit cultural learning. We present a model of the ability to perform inferences about the shared beliefs that underwrite social norms and patterned cultural practices derived from first principles. In helping to solve the puzzle of the *implicit* acquisition of culture, our model provides an integrative view of what has variously been called *mindreading*, perspective-taking, joint intentionality, *folk psychology*, *mentalizing*, or *Theory of Mind* (TOM) — in short, the human ability to ascribe mental states, intentions, and feelings to other human agents and to oneself. To simplify, we will use the term TOM to refer to this ability. Of pertinence to our argument here, TOM (in its various theoretical formulations) is generally described as a key mechanism underwriting the human capacity to form joint goals leading to cultural forms of life (Tomasello, 2009)

As a generative framework TOM has been the subject of sometimes fierce and still ongoing debate in cognitive science (Michael, Christensen, and Overgaard 2014); for a comprehensive review, see (Heyes and Frith 2014). Historically, much of the debate has occurred between

three camps which have advanced alternative explanations for the human ability to infer the mental states of others, namely the Theory Theory (TT), Simulation Theory (ST), and Embodied Cognition (EC) accounts.

Whether one considers the debate settled depends on one's disciplinary and theoretical position. Outside of the field of developmental psychology, which seems to have adopted some arguments from embodied cognition in favour of an enriched TT account, philosophers in the enactivist camp – and to different extents, anthropologists – still disagree with the mainstream 'cognitivist' psychological account of TOM.

Revisiting the TOM debate from the perspective of cognitive and evolutionary anthropology is helpful to contextualise current critiques – e.g., Christensen and Michael (2016); Michael, Christensen, and Overgaard (2014). These critiques stress the importance of considering culture-specific, embodied, and shared interactions with the environment, over the manipulation of internal representations about other minds (reviewed below). Beyond extending debates in the philosophy of mind, the arguments here will be helpful to anthropologists – who are today, due in part to the popularity of the so-called 'ontological turn', e.g., De Castro (2009) – largely committed to anti-cognitivist accounts and psychologists – who largely fail to consider the extent to which cognition is 'collective'.

The basic idea behind TT is that human agents acquire knowledge about the ways in which mental states should be ascribed, which takes the form of a (literal) theory of how minds operate (Gopnik and Wellman 2012; Carruthers and Smith 1996). Proponents of TT hold that social coordination and social cognition require the capacity to make inferences about other people's mental states and propositional attitudes *as such* — that is, an ability to explicitly formulate to oneself that others also think 'silently', that they may hold beliefs that are true or

false, and that there may be a difference between their stated and true intentions, beliefs, or needs; the ability, in other words, to hold a folk theory about other people's minds.

According to a large body of related critiques in the social sciences and phenomenological philosophy, the TT account fails to describe a species-wide mechanism on several counts:

1. TT is a construct derived from Western contexts and fails to describe universal human mechanisms -- we call this the *cross-cultural critique*;
2. TT is a dualistic cognitivist construct, and thus fails to account for the embodied nature of cognition -- we call this the *embodiment critique*;
3. TT is committed to a Machiavellian view of the evolution of cognition that fails to account for the cooperative nature of cognition and behaviour -- we call this the *cooperativity critique*.

The cross-cultural critique

For many anthropologists, the TT account reflects a culture-bound, historically specific notion of 'mind' and the person that is biased toward individualistic Western folk models popularized by enlightenment philosophers (e.g., Locke's notion of personhood as psychological interiority, Cartesian mind-body dualism, Kant's notion of phenomenal reality and selfhood). Critics in this camp point out that many non-Western cultures, folk reasoning about human action does not emphasize individuals' intentions or mental states (Rosaldo 1982; Keane 2015; Duranti 2015; Astuti and Bloch 2015; Luhrmann 2011; Geertz 1973).

Instead, actions may be explained in terms of their perlocutionary effects; that is, in terms of their purported consequences according to locally relevant norms, such as "what would upset the ancestors" (Astuti and Bloch 2015). Extreme versions of this claim have pointed to

ethnographic examples from a group of primarily Melanesian cultures described as having a folk psychology characterized by an 'Opacity of Mind' in which the notion of mental states and psychological interiority is reportedly absent (Ramsey 2007; Robbins and Rumsey 2008).

Recent reviews of this controversy, however, noted that there is no experimental evidence to verify whether and how Melanesians make inferences about others' mental states based on others' behaviour (Robbins, Cassaniti, and Luhrmann 2011), while a close reading of the ethnographic record suggests that folk notions of opacity are *normative* rather than descriptive. This is suggested by ethnographic reports of children being reprimanded for overt curiosity about others' actions or intentions. On this view, Melanesians are simply taught that they *ought not to* wonder about what people are thinking (Robbins and Rumsey 2008; Robbins 2008; Rumsey 2013). Moreover, reports from other Melanesian contexts indicate that it is widely recognized that people 'think silently' (e.g., in the context courtship among the Korowai of New Guinea (Stasch 2009; Luhrmann 2011).

While the current balance of evidence does not support critiques that TT describes a process that is exclusively found in Western cultural contexts, ethnographic studies document wide variation in the ways that people inquire into and talk about others' states of mind that must be accommodated by any account of TOM.

The embodiment critique

Philosophers and psychologists in the embodied cognition camp have also objected to the TT account on the grounds that understanding others or responding to social cues is characterized by 'quick', 'intuitive', 'embodied' responses that need not entail interpretations about other minds or any notion of mental states (Michael, Christensen, and Overgaard 2014). Some of these critics of TT have proposed an alternative approach based on the idea that, rather than mobilizing an explicit theory of mind to ascribe mental states of their, human agents use their

own intuitive, responses to others to understand other human agents and indeed themselves through a process of simulation (Goldman 2006). On the view of such simulation theories (ST), TOM abilities involve processes of modelling others' actions, which may be embodied and automatic (Gallese and Goldman 1998). Embodied cognition need not involve anything that looks like a theory since it uses bodily sensorimotor systems to provide analogical models of human motivation, intention, and action (Shapiro 2010).

Radical enactivist cognitive science takes this emphasis on embodied cognition further to argue that basic cognition does not entail any kind of mental content — particularly not about others' mental states and propositional attitudes (Hutto and Myin 2013). In more recent accounts (Hutto and Myin 2017; Hutto and Satne 2015) enactivists grant the existence of explicit inferences about others, but only in situations that are developmentally contingent on language. Learning to make explicit ascriptions is then a separate, later, developmentally achieved, result of narrative practices (Hutto 2012).

As Heyes and Frith (Heyes and Frith 2014) point out, some current accounts have adopted a compromise position, which gives credence to both sides of the debate, through recognizing multiple processes and progressive elaboration over development. In Apperly and Butterfill's (Apperly and Butterfill 2009) two-systems model, for example, most social cognition may be largely automatic, while a process akin to TT may underpin specific types of language-dependent inferences. Butterfill and Apperly's account stemmed from a growing consensus in cognitive science – famously exemplified in Daniel Kahneman's *Thinking Fast and Slow* (2011) – that cognition can be divided into two "systems": one evolutionarily old, innate, implicit, 'cheap' automatic system of informational foraging supported by a series of largely social biases, and a developmentally-older, evolutionarily young, effortful, relatively inefficient modality of volitional, voluntary reflection. Butterfill and Apperly proposed that the distinction between TT and ST could be cast along this spectrum, with explicit

mentalizing about others entailing a situationally specific, relatively rare sort of reflexivity acquired later in developmental.

Others still have proposed a 'multi-system', progressive scaffolding of socio-cognitive inferences ranging from the fully automatic to the effortfully explicit (Michael, Christensen, and Overgaard 2014). These later 'interactionist' models offer a more nuanced and dynamic account of the gradients of inferences which, rather than being 'located' in discrete cognitive systems, likely occur on a continuum of attunement to different statistical regularities. This is a point elaborated on in detail in Hugo Mercier and Dan Sperber's *Enigma of Reason* (2017), in which they also recast so-called "System 2" reflexivity as varieties of automatic inference *about other's inferences* triggered by communicative cues – actual or imaginary (e.g., in engaging in, or mentally rehearsing conversation and interaction with others). Crucially, these recent models (two systems, multi-systems, interactionist) all study the manner in which agents optimise the metabolic cost of cognition by tuning attentional preference to different domains of statistical regularities, emphasising the function of social and cultural modulations of automaticity. These models, as we argue in 1.3. below, lend themselves to a culturally-informed FEP model.

The cooperativity critique

TOM has played a key role in evolutionary psychology. Early accounts of evolutionary psychology described the evolution of human intelligence and TOM abilities by appealing to the so-called "Machiavellian Intelligence Hypothesis" (Pinker 1999; Trivers 2000; Dunbar 2003; Gavrilets and Vose 2006). On this view, the ability to correctly infer others' mental states — human mindreading — and propositional attitudes about others' mental states evolved through a cognitive arms-race between cheaters (who need to understand others so as to deceive them) and cheater-detectors (who need to understand others to detect deception).

In contrast, scholars in the mutualist camp (Tomasello 2014; Henrich 2015) contend that individual human fitness is best maximized by cooperation with others, leading to an evolved preference for promoting group fitness through the cooperative division of labour. Such cooperation requires knowledge of others' states of mind or intentions. In support of these views, natural pedagogy (Csibra and Gergely 2009, 2011), interactionist (Mercier and Sperber 2017b), and other cultural intelligence paradigms have emphasized the evolved propensity for a non-Machiavellian, cooperative division of cognitive labour, in which mindreading evolved for the purpose of *outsourcing contextually-relevant information* to specific others from our ingroups and to leverage knowledge, skills, and attitudes from a cumulative cultural repertoire. In more radical versions of mutualist models, such as Hrdy's cooperative breeding hypothesis (Burkart, Hrdy, and Van Schaik 2009; Hrdy 2011), mindreading is thought to have evolved in the pre-Sapiens lineage as a result of a 'cuteness and care' arms-race, because selection favoured individuals who were, at once, good caregivers and good at eliciting care from others.

Heyes and Frith (Heyes and Frith 2014) have proposed an account of the cultural co-evolutionary elaboration of TOM abilities, suggesting that the internalist, brain-centred accounts provided by proponents of TT and ST needs to be augmented by an account of how cultural evolution and cultural inheritance sculpt an innate mindreading 'start-up kit', in ways that are analogous to how cultural practices of reading harnessed an evolutionarily older linguistic 'start-up kit' (Dehaene and Cohen 2007).

The extent to which the evolution of perspective-taking abilities requires mental content about other minds is still hotly debated. In the *mindshaping* hypothesis (Mameli 2001; Zawidzki 2008; Zawidzki 2013), for example, mindreading likely emerges from an evolutionarily older and developmentally earlier capacity to imitate, learn, teach, and directly influence others.

Nevertheless, current work suggests that the ability to engage with others as agents with interior states and intentions is central to the cooperative forms of social life we call "culture".

1.3. Piecing together the puzzle of implicit learning: A new portrait of TOM

Conceptualisation

The cultural, embodiment and cooperative critiques of TOM emphasise either *internal* cognitive processes of theory building or simulation or *external*, social-cultural processes of interaction and cooperation. Clearly, these are differences in emphasis and a more complete picture must show how they fit together.

In this paper, we complete this picture by proposing a model of implicit cultural learning that we call "Thinking through Other Minds" (TTOM). In recognizing the virtues (and limitations) of both internalist and externalist accounts, the TTOM model proposes a resolution of the dialectic — and false dichotomy — between so-called internalist (TT and ST) and externalist (mutualist, interactionist, cultural evolutionist) positions.

TTOM integrates a number of recent approaches to the study of cognition, in particular: the cultural intelligence hypothesis in evolutionary anthropology (Tomasello 2014; Henrich 2015; Boyer 2018); the niche construction perspective in evolutionary biology (Laland et al. 2015; Odling-Smee, Laland, and Feldman 2003); the interactionist approach to the evolution of reasoning in cognitive science (Mercier and Sperber 2017b); and the sociocultural enactivist approach to mindreading (Hutto 2012; Gallagher and Allen 2016; Gallagher 2017; Fabry 2017; Hutto, Kirchhoff, and Myin 2014).

What the variational model affords

At a formal level, we integrate these approaches within the framework of the variational free energy principle (FEP) (Friston 2010, 2005) in theoretical neuroscience and biology. Framing this integration in terms of the FEP allows us to derive, from first principles, an interactional model that can explain the acquisition, production, and stabilisation of cultural expectations (Friston and Stephan 2007; Friston 2013; Ramstead, Badcock, and Friston 2017). See Box 2.

We will argue from the formal perspective of embodied (i.e., active) inference, which rests upon our species' remarkable capacity to infer or assign conspecifics to some pragmatic (i.e., prosocial) categories. A successful inference about the 'sort of person you are' enables a host of conditional inferences, many of which have a direct bearing on 'how I should behave'. This is particularly true if I infer that 'you are like me'. We will unpack this view with a special focus on epistemic action, via the selective patterning of salience and attention – and how this is mediated via cultural affordances. We hope to show that these epistemic resources arise naturally from cultural niche construction when, and only when, I share an environment with other 'creatures like me'.

The formalism of the FEP allows us to take further steps toward operationalizing the process of implicit cultural learning and mindreading that we describe as Thinking through Other Minds (see Box 2). In brief, the set of equations that model the process of TTOM could be implemented in computational models, to study simulations of (e.g.) psychophysical, neuronal, and behavioural measurements of the processes involved in a mind-reading or cultural learning task.

On the one hand, such simulations would allow researcher to generate hypotheses about mind reading and cultural learning that may be tested with other empirical methods. On the other hand, FEP simulations can be employed to replicate *in vivo* experiments (e.g., Schwartenbeck & Friston, 2016; Kiebel & Friston, 2011). One can then use the model to explore the dynamic

consequences of changes in parameters associated with the causal factors that led to the generation of the experimental outcomes that were studied empirically. With this method, one also might identify potential contributors to pathological and healthy responses to the task by manipulating the parameters and generating new simulated psychophysical, neural, and behavioural measurements based on the model that has been fitted with *in vivo* data (e.g., Cullen, Davey et al. 2018).

Outline of the argument

Section 2 of this paper introduces the notions of expectations and cultural affordances. We describe *shared attention* and *evolved attentional biases* as crucial mechanisms for engaging with and stabilizing sociocultural niches. We describe the *selective patterning of salience and attention* as the main process behind enculturation, which in turn enables the engagement of human agents with the sets of possible actions (or cultural affordances) that make up their local world (Ramstead, Veissière, and Kirmayer 2016).

Section 3 presents our solution to the puzzle of implicit cultural learning. Human beings acquire the shared habits, norms, and expectations that constitute their culture through their immersive engagement within specific cultural practices, we call "*regimes of attention"* (Veissière 2016). Regimes of attention mark off certain contextually adequate actions as especially salient, and help agents learn to respond to the norms and resources of their local cultural niche. The most important of these resources are the *epistemic resources* that indicate salient information deemed relevant and reliable (Bertolotti and Magnani 2017; Pinker 2003; Clark 2006; Whiten and Erdal 2012).

As we elaborate through the notion of *epistemic authority,* we show that humans are typically biased toward the *source* rather than the *content* of information (Mercier and Sperber 2017b). As amply documented in the literature on so-called 'cognitive errors' (Kahneman 2011), this

tendency can also direct humans toward low-quality, but otherwise high-fidelity information, particularly when it can be intuitively associated with social proof and other mechanisms of social influence (Cialdini and Goldstein 2004). We identify the prestige bias in particular (Henrich and Gil-White 2001) as a central attentional mechanism in the mediation of salience for humans.

The notion of *salience* understood as expected information gain is a central theme of the FEP (Kaplan and Friston 2018; Parr and Friston 2017b, [a] 2017; Friston et al. 2016). Recent FEP based models of cognition-in-context cast niche construction behaviour as the process whereby organisms 'outsource' the computation of salience to statistical structures of the physical environment. The environmental niche then registers information about salience (what an organism trusts or preferentially attends to for it will lead to information gain).

This information corresponds to epistemic resources of the niche (Constant, Ramstead, et al. 2018; Bruineberg et al. 2018; Constant, Bervoets, et al. 2018). Niche construction allows the scaffolding of complex networks of shared expectations encoded across brains, bodies, constructed environments, and other agents, which modulate attention, guide action, and entail the learning of patterned behaviours. Human niches are fundamentally social and cultural — built and constituted by interactions with other people. In the general human niche or any local sub-niche, behaviour is to a large extent *culturally* patterned. Hence, in addition to (and, as we will argue, often prior to) observable statistical regularities in external states of the world, human behaviour is patterned through expectations about *what other people also expect of the world.* It is this domain of expectations about salience and the process of leveraging these expectations that we call "Thinking through Other Minds" (TTOM).

The processes that make up TTOM extend from the conventionalised, normative behaviour of encultured individual agents (e.g., stopping at a red traffic light), which only in some cases

require making inferences about agents, to cases that require *bona fide* inferences about others' mental states for proper, that is, situationally appropriate, modes of engagement.

Section 4 of this paper shows how TTOM integrates standard TOM approaches to tackle the cultural, embodiment, and cooperative critiques. TTOM argues for a compromise position between '*internalist*', brain-based approaches (e.g., simulation and theory-theory theories), which emphasise the neural machinery in individual humans' brains that is necessary to read other minds, and '*externalist*' approaches (e.g., radical enactive and cultural evolutionary theory). Indeed, one of the main motivations for the FEP is to capture the two-way traffic between the organism and the world, to emphasise both the enactment of shared cultural expectations and norms, and the brain-based cognitive abilities that make such an enactment possible, adaptive, and situationally appropriate. Under the FEP, there is no justification for any strict distinction between *dynamics* (as emphasised by externalists) and *inference* (the focus of internalist models).

The conclusion discusses the implications of this model for future research on enculturation and the cultural shaping of cognition in health and illness.

2. Expectations and cultural affordances

In this section, we show that human agents learn most of their expectations through the selective patterning of attention, based on immersive participation in cultural practices. At the outset, we should define what we mean by 'expectations'[2]. We use the term to describe a rich *repertoire* or *spectrum* of priors or beliefs that reflect action-readiness, which ranges from the fully automatic to the effortfully deliberate. Our concept of expectation describes the patterns of action-readiness that modulate and direct the adaptive action of agents; it is thus very broad in its applicability, and ranges from the implicit, embodied expectations that we enact

continuously, often without noticing, to the more consciously held, effortful, psychologically contentful expectations that characterise encultured human consciousness.

2.1. The concept of expectation

On the more automatic end of the spectrum, we can speak of expectations when one's stomach prepares a digestive response upon expecting that food is coming from mastication, or when one's hand and arm prepare an adequate muscle response to lift a half-full glass of wine. Each of these processes reflect different kinds or levels of prior engagement of the world, across different timescales which include evolutionarily old dispositions common to all vertebrates which have been exapted for new uses, as well as distinctive developmental experiences, and learning histories. Together, these elicit physiological, bodily and emotional orientations toward the possibilities for action available in a specific context. Immersion in cultural contexts, moreover, will *structure* such low-level expectations *through participation in patterned cultural practices*; e.g., contextually-patterned modes of affect associated with specific kinds of food and drink, and ritual contexts of consumption.

Human expectations, thus, are always scaffolded through 'levels' (or scales) of evolutionary and developmentally inscribed prior dispositions that come to be modulated by higher-level symbolic conventions (Kirmayer and Ramstead 2017). The intuitive distrust of other people symbolically marked as belonging to an outgroup, for example, has been shown to recruit evolutionarily old disgust responses (Rozin, Haidt, and Fincher 2009; Phillips et al. 1997; Tybur et al. 2013). This involves another level of implicit 'expectations' in which evolutionarily old threat and poison-detection dispositions are activated by (differently implicit) symbolic conventions or affordances (more on which below).

At the other end of the spectrum, many of the expectations that guide behaviour are explicitly taught, effortfully learned, and can be reflected upon (e.g., "sit up straight", "don't fidget in

class"). Such expectations, however, are also more difficult to learn, and least likely to become fully patterned. Indeed, one may sit badly most of the time, fidget in class despite my embarrassment, and face disappointment when one's daughter chooses to become an engineer. Later developing forms of explicit inference require abstract thought, formal instruction, and perhaps deliberation to learn; but once the agent is properly enculturated, new practices usually can be figured out without the direct presence or instruction of other agents. The learner learns the meta-cognitive strategy of how to access, offload, and work with conventional forms of presented cultural knowledge (Heyes 2018). This process, however, will generally entail different modes of indirect social learning e.g., from instructional codes devised by others (like learning a cooking skill from a written guide or YouTube video).

Examining these processes of acquiring conventional or normative behaviours, social scientists have pointed to the important difference between *dogma* (official doctrine) and *doxa* (common belief) (Bourdieu 1977). The explicit rules and conventions established in dogma (what people know they must do) and reported in everyday speech are poor indicators of the regularities of a culture – and how humans learn cultural behaviour in general. *Doxa*, in Pierre Bourdieu's famous formulation, refers to *all that is taken for granted* in any given context or society. For instance, in his 'dramaturgical' account of social life, sociologist Erving Goffman (Goffman 2009) describes the gradients of effort and explicit performance required in the obedience to and enactment of social conventions in everyday life. Goffman notes that in some spaces (like the home), which are symbolically marked as the 'backstage', people tend to relax their effortful behaviour and ignore or disobey many social rules; they trade off the dogma for the doxa. Nevertheless, their behaviours necessarily draw from the culturally shaped repertoire of normative and conventional forms.

What interests us here is how the doxa of backstage behaviour (indeed most of solitary cognition) is itself already culturally patterned, despite the immediate absence of others'

enforcing gaze (and the foregrounding of inferences we make about what others know and expect in context). A first hint is the fact that human agents are constantly (deliberately or automatically) adjusting what they are doing to what relevant others (e.g., role models or anti-role models, specific or generalized) *expect*, and *expect them to expect*, and so on. Much of this is accomplished implicitly (Tomasello et al. 2005); usually through nonverbal communication with gesture, facial expression, posture, and pantomime, but also through language when necessary. Evidence that this kind of expectation does not depend on language comes from the observation that infants as young as 15 months are able to make *implicit inferences* about others' mental states (Onishi and Baillargeon 2005) and actions well before they can formulate *explicit statements* to this effect (Michael, Christensen, and Overgaard 2014).

2.2. The concept of affordance

In Gibson's ecological approach to perception (Gibson 1979), things and features of the world are said to *afford* possibilities for engagement (Chemero 2009; van Dijk and Rietveld 2016). An affordance is a relation between an agent's abilities and the physical states of its environment. For instance, water affords drinking, cups afford drinking-out-of, bridges afford crossing, axes cutting, handles holding, etc. Affordances are defined in terms of physical properties of the thing in the world (e.g., being graspable, being able to support the weight of a person) and in terms of the *abilities* or expectations of the agent (e.g., knowing how to sit straight). Abilities can be described in terms of the spectrum of *expectations* with which the agent is endowed (Gibson 1979; Pezzulo and Cisek 2016; Rietveld and Kiverstein 2014; Tschacher and Haken 2007). It takes an agent with a mouth, throat, stomach, etc. (to drink), and hands and opposable thumbs (to grasp a cup), and a certain set of skills (hand-eye coordination, for instance) to be able to 'discover' the relationship of water and cups to the action of drinking.

The relation of affordances to the notion of expectations is a recent extension of the ecological approach that explains perception as conditioned on the beliefs of the agent (Bruineberg and Rietveld 2014; Chemero 2009). Hence, affordances are not simply static features of the environment, independent of the presence and engagement of an agent; nor are they states of the cognitive agent alone. Affordances are "invariant variables" or structures of relatedness (Gibson 1979), (Gibson 1979 p.134). In the case of sensorimotor affordances, for example, they are invariant, in that they are grounded in the physics and geometry of the agent's interaction with the environment, which results in relationships that are highly reliable and stable across time, and are ready to be perceived or (re)discovered by the agent; and they are variable, in that they are specified dynamically by the sensorimotor and other cognitive abilities of the agent. In the case of affective affordances and expectations, the stability may reside in the neurobiology of organisms' learning and memory systems coupled with the persistence of the environmental cues to which particular patterns of recollection and enactment have become linked. The relational space of possibilities between agents and their environments constitutes an *ecological niche*. Agents and their environments are modified, and become attuned to each other, as the result of their history of co-adaptive interactions (Bruineberg and Rietveld 2014; Gibson 1979).

These examples are congruent with work on the evolution and cultural learning of tool use (Stout et al. 2008; Stout and Chaminade 2007), which illustrates the need for humans to learn to hierarchically structure actions with long-term consequences. 'Hierarchical' here means that actions are nested within one another, and that complex behaviours require planning a whole chain of nested actions, not just the immediate optimization of current actions or a simple sequence. This kind of executive control of behaviours is characteristic of *enculturation*, in which complex sequences of action are built out of iterative structures of simpler components strung together in ways that reflect the results of collective experiences

of trial and error. An individual is therefore able to borrow from and integrate the experimentation and learning of others in the cultural group.

Direct or 'natural' affordances in the humanly-constructed ("anthropogenic") environment can be supplemented, modified or supplanted by 'conventional' affordances (Ramstead, Veissière, and Kirmayer 2016), which depend on shared cultural conventions, based on skills learned through immersive social practices. Thus, bodies of water ('naturally') afford drowning for all humans, and swimming for those with the acquired skills that allow them access to that specific *cultural* affordance. Mastering swimming, like all cultural affordances and most of what humans do and think, requires immersive participation (Roepstorff, Niewöhner, and Beck 2010; Hutto 2012), which includes imitation, practice, repetition, and a grasp of norms and conventions. Thus, affordances are contextually sensitive. For example, for the right kind of agent, a formal suit and tie might function as a cue that indicates authority and affords deference; but when additional cues are added (e.g., a napkin draped over the forearm and a silver tray with glasses), the affordances will change whose enculturation enables them to respond appropriately to the cues.

2.3. Learning cultural affordances

How are the affordances of the niche learned? What does it mean to learn to recognize and engage a specific field of affordances? This is a puzzle, since affordance theory tends to collapse basic categories of learning like 'knowing how' and 'knowing that'. For instance, there is no necessary precedence of the knowing 'that' a cup is for drinking over the knowing of 'how' to drink from a cup, and vice versa. Even in domains where knowing 'that' seems to precede knowing 'how', such a distinction does not hold, since knowing 'that' is leveraged as a skill interiorized and integrated to normal implicit motor practice; e.g., architectural design (Rietveld and Brouwers 2017) and mathematical thinking (Menary 2010). Put simply,

knowing ‘that’ is only knowing ‘that’ when it becomes know ‘how’, and acquiring know ‘how’ requires interiorizing and embodying know ‘that’. This circularity can be understood through a process of scaffolding that occurs on multiple temporal scales associated with: the cultural co-evolution of particular niches, communities or traditions; the developmental trajectory of individuals; and the process of learning to engage with new social contexts.

What, then, are the underpinnings of scaffolding? Some anthropologists, like Tim Ingold, have argued that human niches comprise affordances that can be figured-out, rediscovered, or rebuilt by human individuals in each generation without the ‘transmission’ of a purportedly separate realm of ‘cultural representations’ (Ingold 2001). Critics of Ingold, e.g., (Howes 2011) have pointed out that most of what humans learn over their life spans in order to become proficient at functioning in their local worlds, is learned *socially* – that is to say, learned primarily from other humans, and not just from what things or situations themselves afford. However, Ingold maintains that many aspects of human life are simply emulate (Hamilton 2008), ‘shown’, or ‘pointed to’, and left to be explored, ‘figured out’ and experimented with by individual learners (for example, in play).

The main role of others in this kind of social learning is to direct attention rather than to convey specific semantic content (Tomasello 2014). In effect, social learning involves immersion in local contexts through what we call *regimes of attention* and imitation that direct human agents to engage differentially in forms of shared intentionality. We have argued that such regimes of attention play a central role in the enculturation of human agents (Ramstead, Veissière, and Kirmayer 2016). Indeed, human beings seem particularly specialized for such forms of social learning (Sterelny 2012).

Humans mostly learn deictically (in context) and pragmatically by participating in cultural practices and by being immersed in the ways of doing things that characterize a given local

culture. Some of this involves following the "tracks" laid down in local environments by others, or following the norms and rules presented through by institutions, without engaging with others' interiority. But many convention-dependent forms of learning require inferences based on prior knowledge about how we expect others to think and behave in specific settings (e.g., adjusting to culturally-specific turn-taking rituals in public space) (Ramstead, Veissière and Kirmayer 2016).

The process of learning how to engage cultural affordances to think through other minds likely begins in infancy when we seek or accept guidance from our caregivers, and further develops through exposure to social hierarchies of prestige, themselves embodied in kinds of high-status agents that can be leveraged as models (Feinman 1982), which are knowledgeable or skilful ingroup members, educators, community and religious leaders, celebrities, and imaginative reconstructions of folk or historical personages with high epistemic prestige (e.g., "What would Wittgenstein think of this theory?"). Individual action, in turn, is guided by what agents expect relevant agents to expect of them ("What would mother expect me to do?").

Others in our social world present us with cultural affordances as well as solicitations for action. Engagement with these realizes a specific social niche, context, group or community. The reliance on social and cultural affordances co-constructed with and maintained by other people makes it important for us to distinguish between those who think like us and those whose thinking is either systematically different from our own or else unfamiliar and, hence, unpredictable – and inherently surprising. This distinction marks off domains of in-group and out-group, with corresponding epistemic authority. Regimes of attention then make the right kinds of social solicitations stand out in context, thereby allowing the learning of socially relevant affordances in a given cultural niche, community or local world.

2.4. The phylogeny and ontogeny of cultural affordances

In human ontogeny, it is likely that affordances are first learned implicitly, automatically, and with little conscious effort, through imitation, repetition, and rewards. Phylogenetically, the human mind evolved to support a series of adaptive 'content biases' (Henrich 2015) for features of the world that possess high intrinsic learnability, and feed-forward potential through teachability and memorability. Fire, edible foods, and simple tools, for example, all have been amply documented as possessing these heuristic properties (Henrich 2015). In the realm of more conventional affordances, compared to other primates, humans are also unusually adept at tracking other agents' social status and shifts in symbolically-assigned prestige through gossip (Dunbar 2004; Henrich and Gil-White 2001).

Status among social animals generally provides a guide for whom to follow and obey, and from whom or what to learn. As cultural evolutionists have pointed out (Mercier and Sperber 2017a; Henrich and Gil-White 2001), social status among humans serves a primarily *epistemic* function. One seeks guides for thought, behaviour, and affect in agents who embody sources of relevant cultural information that are deemed to be of high quality in relevant social contexts (e.g., we learn from professors in the classroom, and seek help from good students, or seek to publish in high impact journals). Among humans, symbolically-conferred *prestige* has largely replaced sheer physical dominance as a way to find, acquire, and signal status (Henrich 2015). In social context, marks of distinctions (Bourdieu 1984) such as styles of dress, forms of speech, and other techniques of the body provide a shortcut that signal an agent's status on the various prestige scales deemed relevant. *Gossip*, in turn, serves the more fine-grained communicative function of keeping track of an agent's conferred prestige and epistemic status.

The mechanisms described above rely on evolved cognitive biases for cultural transmission that have been hypothesised to serve an information-tracking function (Henrich 2015); that is, as enabling humans to outsource their decision making to other *agents*, through patterned *interactions* with them and the *shared places* in which they dwell. The physical structure of the environment – including artefacts, practices, and other socially constructed aspects of the ecological niche – embody or encode adaptive, context-relevant cultural information endowed with *salience* – that is, as high-quality, or 'useful' sources of information in context. A dramatic illustration of this is provided by the infamous Milgram experiments (Milgram 1963), which demonstrated the extent to which human agents are ready to outsource their actions to those that symbolically display the right credentials and wield epistemic authority.

Social status serves the epistemic function of locating the person in a locally relevant hierarchy – a process that can also be described in terms of affordances as prestigious agents solicit imitation through such perceived qualities as *trustworthiness* (Mercier and Sperber 2017b), and *credibility* (Henrich 2015). How well or badly agents respond to such *affordances*—as indexed through gossip, e.g., circulating stories about cheating spouses, embezzling chiefs, or free-riding subordinates —thus will largely determine the levels of trust that they inspire in others. Furthermore, the hierarchy that locates the person is not only material but also symbolic, as expressed through historically acquired and social displayed marks of distinction. This poses a challenge to an account of affordances in terms of immediately present features.

Humans are accustomed to attending to certain people, in certain places for tones of voice, facial expressions, shifts in body posture, etc., which signal approbation, disapproval, or moral concern, and hence convey (in context) normative information (Williams 2011; Ignatow 2009). As we have seen, beyond what they naturally afford, human material environments have additional, *symbolically-inscribed* normative and deontic powers that

deeply permeate the way that individuals affectively approach and engage with their niches (Kaufmann and Clément 2014). For instance, in the European Middle Ages, children may have been socialized to fear forests as dark and dangerous spaces full of beasts, witches, and evil spirits through folktales and bedtime stories. In contrast, in many hunter-gatherer cultures, like the Aka of Central Africa, children are equipped with cultural knowledge to expect the forest to offer a safe, nurturing space (Hewlett 1994; Hewlett 2017).

The physical environments occupied by various human groups and sub-groups also characterises *group-specific* affordances (e.g., a neighbourhood or a city) (Einarsson and Ziemke 2017). Consider how a space (e.g., a university or museum) that is symbolically marked with group-general standards of prestige – a space, thus, that has been historically inaccessible to low-status individuals – will afford radically different experiences to high and low status individuals depending on how their respective subgroup is valorised in their macro-cultural niche. Pierre Bourdieu's concept of *habitus* (as the internalization of social norms in techniques of the body) is one way of approaching the varying effects of a sociocultural niche on individuals with different status or position. To expand on Bourdieu's (Bourdieu 1977) reflections on the effects of cultural capital on *habitus*, we note that a similar space can be marked as 'welcoming' for some, but as 'intimidating' or outright 'hostile' to others (e.g., for minority groups). This reflects a related, orthogonal distinction between the familiar (predictable) versus the unfamiliar (unpredictable). From a cultural affordances perspective, being socially marked and positioned at a particular place in a cultural niche enables automatic responses in one's patterns of movement, posture, breathing, and gaze, as well as in neurobiological responses, such as fluctuations in cortisol (Bijleveld, Scheepers, and Ellemers 2012), oxytocin (Hrdy 2011; Luo et al. 2015), or testosterone (Cheng et al. 2013).

The co-existence of habitus or internal physiological dispositions with external features of an adaptive niche points to a crucial feature of affordance theory, namely, that the affordances of

the environment and the capacities of an individual are inextricably interwoven, and co-determining. However, developmentally, and in shared social contexts, culture precedes individual action and experience. In a sense, culture confers on the environment latent affordances such that, if one learns the right repertoire of skills (including attentional strategies) from one's forebears (by acquiring specific cultural knowledge and practices) one can 'read' the environment in new ways, thereby discovering 'new' affordances (that were, in a sense, there all along, insofar as they engaged other or prior skilled actors). Moreover, since one of the functions of cultural affordances is to allow improvisation (and hence the creation of new cultural forms), the affordances of a niche that are being actively engaged are always in the process of discovery, elaboration and extension. Clarifying the temporal move from group or cooperative affordances to individual ones (and back) is part of explaining developmental enculturation, skill acquisition, and culture-production.

So far, we have described regimes of attention and symbolic layering as cultural affordances of the *conventional* and *normative* variety. Over the course of human ontogeny, this 'conventional' domain of culture eventually becomes superordinate to the natural domain. Past a certain developmental stage, language can be used to install superordinate frames through which subsequent affordances are perceived and engaged (cf. Bengio 2014). This linguistic capacity to leverage affordances can include cooperative behaviours that reflect social norms and cultural forms of life. The statistical regularities exploited in learning cultural affordances, thus, are primarily situated in the realm of *expectations that humans learn to form about other people in the niche*; that is, in the realm of *folk psychology*. We call this intersubjective process of engaging others' expectations and inferences "*Thinking through Other Minds*" (TTOM). In the next section, drawing on the FEP, we turn to the question of how cultural affordances can be acquired and maintained to coordinate large cultural groups, through selective patterns of attention and learning.

3. TTOM: Learning Cultural Affordances Under the Free Energy Principle

3.1. The free-energy principle as applied to individual cognition

To explain cultural affordances and implicit cultural learning, we draw on the variational free-energy principle (FEP). The FEP is a mathematical statement of the fact that living systems act to limit the repertoire of physiological (interoceptive) and perceptual (exteroceptive) states in which they can find themselves (Friston, Kilner, and Harrison 2006; Friston 2013). Although even simple organisms have autoregulatory mechanisms to restrict themselves to a limited number of sensory states (compatible with their survival), humans additionally accomplish this feat by leveraging cognitive functions and socioculturally installed behaviour. For instance, if core body temperature drops from its usual 37 degrees Celsius, internal processes of shivering are automatically evoked and externally oriented actions are initiated to move the agent toward a heat source, or to put on a jacket or parka.

This requires the agent to learn about the structure of its environment, which, from the point of view of the brain, is not a small business, since the (skull-bound) brain is secluded from the causal regularities in the environment it seeks to learn (Hohwy 2013).

The brain only has direct access to the way its sensory states fluctuate (i.e., sensory input), and not the causes of those inputs, which it must learn to guide adaptive action (Clark 2013) – where ‘adaptive’ action solicits familiar, unsurprising (interoceptive and exteroceptive) sensations from the world. The brain overcomes this problematic seclusion by matching the statistical organization of its states to the statistical structure of causal regularities in the world. To do so, the brain needs to re-shape itself, self-organizing so as to expect, and be ready to respond with effective action to patterned changes in its sensory states that correspond to adaptively relevant changes ‘out there’ in the world (Bruineberg and Rietveld

2014). Because action selection and response conforms to such expectations, behaviour can effectively maintain the agent within expected states.

The FEP describes this complex adaptive learning process in terms of variational inference (also called approximate Bayesian inference). Briefly, the idea is that the agent learns a statistical model of sensory causes in the world, called a *generative model*. This model represents the agent's relation to the environment, and enables it to predict how sensory inputs are generated, by modelling their causes (including, crucially, the actions of the agent itself).

The generative model underwrites the agent's perception and action as they unfold over time. The parameters of the generative model encode the beliefs of the agent about its relation to the environment (e.g., When I move my finger to flip the switch, the light goes off). This is realised by neural network dynamics that change over short timescales (reflecting external states of the world), and slower changes in network connectivity that encode parameters that change over longer time scales to reflect the contingencies that underlie the agent's representations of the transitions among the states of the world (e.g., the probability of my finger's moving the switch to change its state from 'down/off' to 'up/on') (Kiebel, Daunizeau, and Friston 2008).

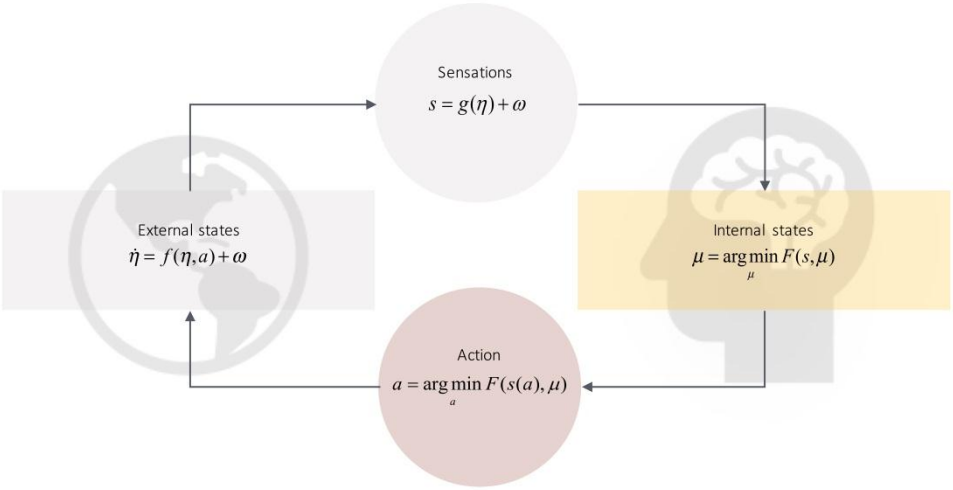
The generative model functions as a point of reference in a cyclical (action-perception) process that allows the organism to engage in active inference. Internal states of the agent (e.g., the states of its brain) encode a *recognition* density; that is, a probability distribution or Bayesian belief about the current state of affairs and contingencies causing sensory input. This (posterior) belief is encoded by neuronal activity, synaptic efficacy, and connection strength (Friston 2010). The mathematical formulation behind the FEP claims that all of these internal brain states change in a way to minimise variational free energy. By construction, the variational free energy is always greater than a quantity known as *surprisal*, *self-information*

or, more simply, *surprise* in information theory. This means that minimising free energy minimises *surprise*, which can be quantified as the negative logarithm of the probability that 'a creature like me' would sample 'these sensations'.

Crucially, in minimising free energy, the posterior beliefs encoded by neuronal quantities approximate the true posterior density over the causes of sensations (see Figure 1 for details). Intuitively, the variational principle of least free energy is just a description of systems (like you and me) that seek out expected sensations. An equivalent and complementary interpretation follows from the fact that surprise is the converse of Bayesian model evidence in statistics. This means that we can understand active inference as gathering sensory evidence for an agent's model of its world – sometimes referred to as self-evidencing.

Put another way, this can take the form of seeking expected sensations associated with novelty or danger (e.g., thrill-seeking) or in more maladaptive cases (e.g., depression), of 'confirming' the negative valence of one's world through rumination (Badcock et al. 2017). As we discuss in section 3.3. below, accounting for novelty-seeking in free-energy minimisation is an important contribution of the model. On the face of it, humans seem to find some a certain kind of surprise desirable. To understand this mathematically, it is useful to appreciate that *expected surprise* (i.e., expected free energy) is *uncertainty* (i.e., entropy). This means that certain acts such as 'attending to this' or 'looking over there' become attractive if they afford the opportunity to reduce uncertainty. Think of the game of 'peek-a-boo' played with infants as a case in point, in which the infant (as learned through repeated practice) attends earnestly in pleasurable anticipation of resolving uncertainty about where her mother will reveal herself. Generally speaking, epistemic affordance of this sort has a positive valence because it entails a reduction of uncertainty; both about states of affairs in the world – and 'what will happen if I do that'.

In summary, the FEP – as applied to individual cognition – describes the process by which an agent updates its (Bayesian) beliefs, encoded by brain states, to optimise a generative (in the sense that it makes predictions) model of the world. When these beliefs are realised by action upon the world, this process is known as *active inference* (Friston 2011; Friston, FitzGerald, et al. 2017). Active inference involves the coordination of sensorimotor patterns (i) by selectively sampling sensations that minimise expected surprise (i.e., by actions that include orientation, attention, and exploration) and by (ii) updating expectations about the most probable causes of sensory inputs (i.e., perception). Perception entails optimising beliefs about states of the world and learning the parameters of generative models, via Hebbian processes of associative learning (Friston 2010).



$$F(s,\mu) = \underbrace{E_q\left[-\ln p(\eta,s)\right]}_{Energy} - \underbrace{H\left[q_\mu(\eta)\right]}_{Entropy}$$

Action minimises the bound on surprise

$$F(s,\mu) = \underbrace{D[q_\mu(\eta) \,||\, p(\eta)]}_{Complexity} - \underbrace{E_q[\ln p(s(a) \,|\, \eta)]}_{Accuracy}$$

$$a = \arg\max_{a} Accuracy$$

Perception optimises the bound on surprise

$$F(s,\mu) = \underbrace{D[q_\mu(\eta) \,||\, p(\eta \,|\, s)]}_{KL\ divergence} - \underbrace{\ln p(s)}_{Surprisal}$$

$$\mu = \arg\min_{\mu} Divergence$$

Fig. 1. Self-evidencing and the Bayesian brain: Upper panel: Schematic of the quantities that define an agent and its coupling to the world. These quantities include the internal states of the agent (e.g., a brain) and quantities describing exchange with the world; namely, sensory input and action that changes the way the environment is sampled. The environment is described by equations of motion that specify the dynamics of (hidden) states of the world. Internal states and action both change to minimise free-energy or self-information, which is a function of sensory input and a probabilistic belief encoded by the internal states. Lower panel: Alternative expressions for free-energy illustrating what its minimisation entails. For action, free-energy (i.e. self-information) can only be suppressed by increasing the accuracy of sensory data (i.e., selectively sampling data that are predicted). Conversely, optimising internal states make the representation an approximate conditional density on the causes of sensory input (by minimising a Kullback-Leibler divergence between the approximate and true posterior density). This optimisation makes the free-energy bound on self-information tighter and enables action to avoid surprising sensations (because the divergence can never be less than zero). When selecting actions that minimise the expected free energy, the expected divergence becomes (negative) epistemic value or salience, while the expected surprise becomes (negative) extrinsic value; namely, the expected likelihood that prior preferences will be realised following an action. Please see Appendix for a technical explanation – and description of the variables in this figure.

3.2. Attention and learning

Not all kinds of sensory inputs are equal in their significance or reliability, and therefore, they need to be differentially weighted when updating beliefs via free energy minimisation. For example, interoceptive signals might merely be tracking physiological noise (Seth and Friston 2016; Feldman 2013), or again, exteroceptive sensory streams can stem from anomalous events that are unlikely to recur. Nevertheless, *a priori*, any signal can indicate relevant

information that is worth accumulating, insofar as it enables an agent to track statistical regularities of the niche. An important aspect of self-evidencing involves updating beliefs about the reliability or precision of sources of information, particularly, sensory input. *Sensory precision* corresponds to the precision of sensory information; e.g., how much confidence or reliability can be afforded auditory input, when a rabbit listens out for a fox sneaking in the grass.

Since the agent has to navigate a capricious and context-sensitive environment, it also needs to assess the *precision of its own expectations*; namely, how far expectations depart from typical beliefs. This corresponds to *prior precision*; e.g., how much confidence or precision a rabbit should afford its prior beliefs, *given its expectation*s about the presence of foxes in the area at that time of the day. Note the subtle but fundamental difference between expectations or beliefs about the (first-order) causes of sensations and expectations about precision, which constitute (second-order) estimates of statistical context (Hohwy 2013). In short, precision reflects the reliability of expectations about states of affairs; i.e., whether or not sensory evidence or prior beliefs can be *trusted* (and not what they concern *per se*).

Using the FEP, we can distinguish two complementary, but computationally distinct, aspects of the folk-psychological concept of 'attention' (Parr and Friston 2017a, 2018, [b] 2017): (1) as the process of directing the organism to selective sampling of the world (through shifting attention, sensory modulation, movement, or exploratory behaviour) such as to resolve uncertainty (i.e., expected surprise)[3]; and (2) as the calibration or weighting of this information as it is gathered to minimise surprise. Both play a crucial role in what follows. Under the FEP, *salience* is considered the main candidate for the implementation of attentional processes in the first sense; namely, the information gain or resolution of uncertainty afforded by the active sampling of the sensorium. The second sort of attentional selection corresponds to *precision-weighting* (the modulation of belief updating as a function

of estimated precision). This attentional process selects certain (neuronal) messages for belief updating through differential selection or modulation (Stephan et al. 2008). In short, *salience* is an attribute of action – in the sense of a particular way of sampling the world has epistemic affordance, while attentional *selection* via precision weighting is an attribute of perception – in the sense of accumulating the right sort of information after it has been sampled.

Figure 2 illustrates the attentional selection of messages using a predictive coding formulation of free energy minimisation. In this formulation, prediction errors are passed upward through hierarchical connectivity architectures in the brain to update higher order expectations. In turn, the expectations provide descending predictions to create prediction errors. In this scheme, sensory precision is assigned to prediction errors at the sensory level of the hierarchy, while prior precision is assigned to prediction errors at higher levels. This precision weighting is thought to underwrite attentional selection of sensory input and is a crucial aspect of perceptual inference (Feldman and Friston 2010; Hohwy 2013). In what follows, we will subsume both sorts of attentional mechanisms under salience, given that overt sampling and covert attentional selection[5] both conform to the same variational principles, under the FEP.

Attentional salience plays a central role in learning to engage with culturally constructed niches, both to select sensory evidence relative to the individual's goals and to identify sources with high reliability. The cultural affordances model proposes that human agents acquire culture by being immersed in specific, culturally patterned practices that modulate salience, which we call 'regimes of attention' (Veissière 2016; Ramstead, Veissière, and Kirmayer 2016). Most regimes of attention do not involve isolated independent features of the environment, but correlated cues and opportunities for epistemic action that are organized in terms of local, cultural forms of cooperative activity, norms, and practices.

As we will describe in section 3.4., and as shown in Figure 3, these epistemic actions are supported by epistemic resources offered by the local cultural niche. In turn, regimes of attention correspond to the *salience* or epistemic affordance of sources of cultural information embodied in the epistemic cues of the niche. As shown in Figure 2, through active inference over the local cultural niche, humans can learn the norms and other contingencies that govern their local cultures.

Crucially, the configuration of regimes of attention by cultural practices and the ensuing attribution of salience to cultural information is only one of two aspects of cultural learning under active inference. The other aspect is the modulation of salience *via the modification of the environmental aspects of the patterned cultural practices* (e.g., people and material artefacts). As we will see in section 3.4., this 'external' modulation of salience is enabled by mechanisms that we associate with developmental niche construction broadly construed (by analogy to internal mechanisms, such as perception and learning in the brain) (Constant, Ramstead et al 2018; Constant, Bervoets et al 2018; Bruineberg, Rietveld et al. 2018). Indeed, most predictions made by human agents result from — and pertain to — interactions with other human agents that co-construct a shared local culture and its niches. Through these niches, this culture furnishes feedback for the neurocognitive processes that serve the cultural patterning of attention (Seligman, Choudhury, and Kirmayer 2015). As such, it follows that what we call 'culture' is an extensive process that recruits elements both within the brain and in the shared cultural world (e.g., constructed places and designed artefacts).

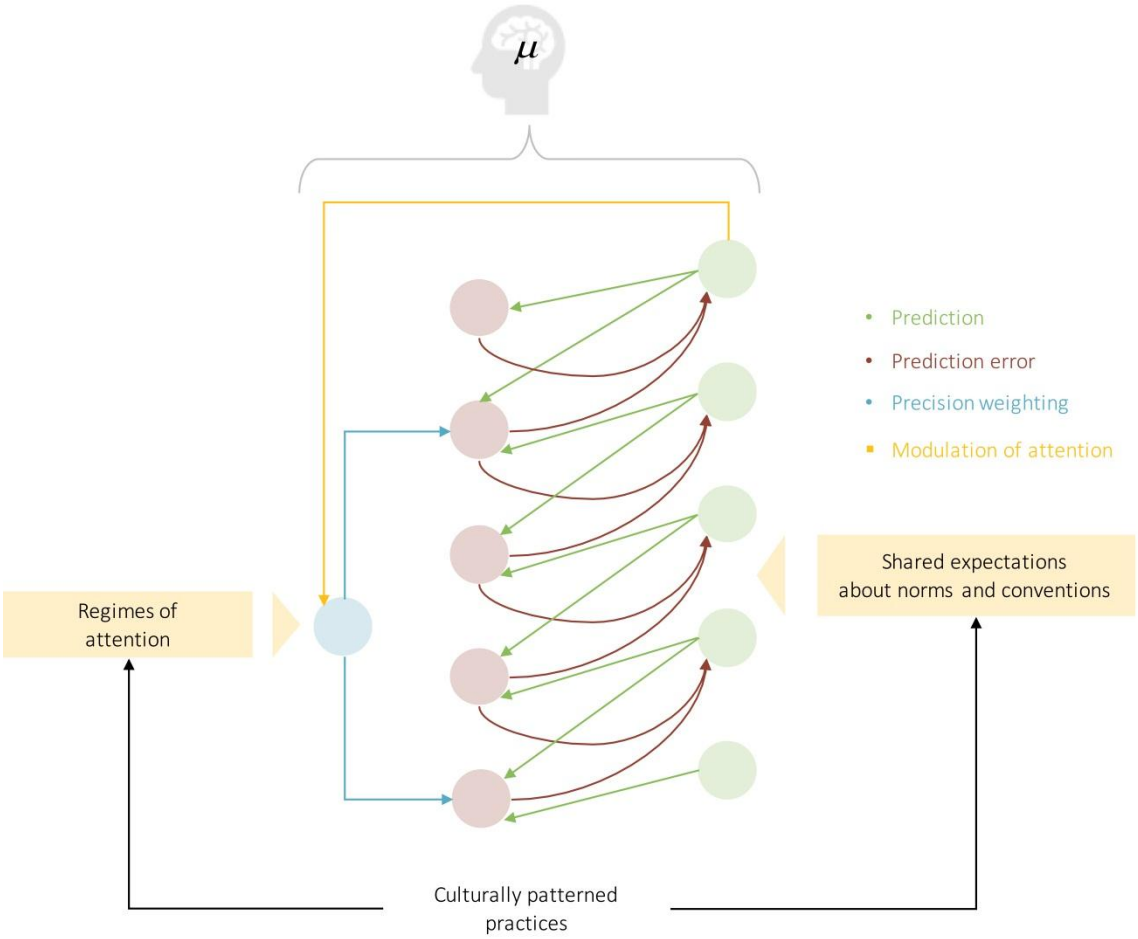


Fig. 2. Cultural affordances. A schematic illustration of the looping effects that modulate social learning by human agents through expectations that, in turn, enable their interaction with cultural affordances. The attentional processes of individual agents are modulated by regimes of attention and by the shared expectations, norms, and conventions that characterize their local culture. In this example, the key point is that the yellow arrows effectively bias self-evidencing towards or away from (certain kinds of) sensory evidence – and that the optimal selection (i.e., salience) has to be both learned and learnable in the right sort of cultural context. Adapted from (Ramstead, Veissière, and Kirmayer 2016)

3.3. Novelty, salience, and surprise

One might argue that there is an important design specification issue here; that is, to what patterns is salience or epistemic affordance attached (e.g., specific sensory information, families of similar events, sources of information)? Any such assignment implies a pre-existing conceptual structure that allows for parsing the flow of information and that imparts some kind of hierarchical organization to available information. Precision and salience estimates are judged against some notion of what is salient (and this cannot just be what is stable over time, since that could result in a small, self-satisficing circle).

Under the FEP, these design specification issues are addressed by assuming that the agent embodies expectations that are established through histories of learning and, ultimately, through natural selection (Friston 2010; Badcock 2012; Badcock et al., 2019). Prior expectations are heritable through genetic, epigenetic, and exogenetic mechanisms (Constant, Ramstead, et al. 2018). These specify the epistemic value of sensations, and by the same token, the extent to which they should be considered. Priors that are inherited by the agent thus mandate the occupation of a limited repertoire of sensory states with high epistemic value that are revisited again and again (Friston et al. 2015; Pezzulo and Cisek 2016; Friston 2010), thus giving the impression that the agent maintains its organization – i.e., limits or minimises the free-energy of its phenotypic states with regard to the states in its niche. Our account thus focuses on the *conservative nature of human culture*; its ability to ensure that certain well-bounded and highly valuable states are frequented.[4]

Conservation is essential to cultural continuity and enculturation, but cultural niches also constantly change through creative innovation and adaptation. This raises the question of how free energy minimization and dynamical coupling can account for creativity and innovation in social coordination, behaviour patterning, and the organization of sociocultural ensembles.

Proponents of the FEP face a similar issue at the level of individual cognition, known as the 'dark room problem' (Friston, Thornton, and Clark 2012; Kiverstein, Miller, and Rietveld 2017). The problem is simple: if agents aim to avoid unexpected encounters with their environment, we should expect minimally changing sensory environments like dark rooms and correspondingly monotonous sensations to be the most frequently (re)visited states of an organism. Yet, there are countless examples in every aspect of life (from art and politics to eroticism, contemplation, and drug-taking, to name but a few) in which humans seem motivated (or driven) to *maximize* novelty, and evanescent states of being (Veissière 2017). What, then, prompts novelty seeking behaviour at the level of individuals and social ensembles?

The FEP deals with the issue of novelty seeking behaviour by formalising action as being in the game of maximising the *epistemic value of action* (or *epistemic affordance*). In essence, free energy minimizing agents seek to sample the world in the most efficient way possible. Since the information gain (i.e., salience) is the amount of uncertainty resolved, it makes good sense for the agent to selectively sample regions of environment with high uncertainty, which will yield the most informative observations. This relates to the development of *artificial curiosity* in neurorobotics as a form of *intrinsic motivation* – so called because the resolution of uncertainty is itself intrinsically valuable and drives exploration (Oudeyer and Kaplan 2007; Schmidhuber 2006; Friston, Lin, et al. 2017; Friston, FitzGerald, et al. 2017).

In effect, agents will act to optimise the *epistemic value or affordance* of an action before acting on its *pragmatic value*, which is essentially its expected utility (Friston et al. 2015; Pezzulo et al. 2016). For example, if one enters a dimly lit kitchen to grab a midnight snack from the pantry, one is more likely to turn the light switch on before heading to the pantry. Turning the light on allows one to get an optimal grip and disambiguate the situation, before one acts on the pragmatic value (i.e., the utility) offered by snack foods. In short, the dark

room objection fails because it simply does not take into account the formal description of action under the free energy principle. In selecting action, an active inference agent (a.k.a. a free energy minimising agent) attributes an intrinsic value to the reduction of uncertainty, which entails exploration. Hence, under active inference, policy selection fundamentally is guided by intrinsic, epistemic (belief-based) imperatives. This formally differentiates approaches based on the FEP from non-epistemic (belief-free) formulations, such as reinforcement learning (Cullen, Davey et al. 2018).

Intrinsic motivation[5] and artificial curiosity enables agent to explore novel, transient, and unexpected regions of the space of policies open to them. This can be an 'adaptive' exploration or epistemic foraging, since it allows for the exploration of this space; over longer timescales, the local increase in free energy serves the more general process of reducing free energy (either for the individual, because it prepares the organism for potential changes in adaptive contexts, and enlarges the repertoire of responses for the individual or the group). Similarly, cultural diversity allows individuals and groups to explore alternative niches that may provide adaptive advantage in the larger fitness landscape (Bengio 2014).

This can be seen on the temporal scale of human cultural co-evolution. The 7R variant of the DRD4 gene (which encodes the D4 subtype of the dopamine receptor) appears to have become more widespread 50,000 years ago at a time of great migrations and a revolution in hunting technology among early *Homo Sapiens* (Andrews, Gangestad, and Matthews 2002; Swanson et al. 2002; Shelley-Tremblay and Rosén 1996). Traits like novelty-seeking, creativity, high energy, and willingness to take risks associated with that gene likely conferred adaptive advantages in the environment of our ancestors. These may have become less valuable or even maladaptive later as human niches became safer, more standardized, and more predictable. Indeed, this shift in adaptive value with cultural context is invoked in evolutionary explanations of some forms of behavioural dysfunction (like Attention Deficit

Hyperactivity Disorder) (Shelley-Tremblay and Rosén 1996; Tovo-Rodrigues et al. 2013). Of course, even maladaptive (non-optimal) traits may come to be culturally valued or exploited by individuals and communities, perhaps to their own detriment. Only the first of these pathways relates to the normal, adaptive acquisition of culture, which is the main focus of this paper. However, both forms of epistemic foraging might contribute to cultural evolution.

3.4. Niche construction and learning

Culturally competent agents must learn regimes of attention across similar kinds of situations. For example, drivers must learn how pedestrians waiting at a red traffic light or crosswalk behave. The norms of pedestrian-vehicle behaviour vary in different cultural contexts. In some local contexts, pedestrians have the right of way and cars must stop, or pedestrians may observe red lights more laxly and attempt to cross against a red light, if the traffic is sparse. Within a given context individuals' behaviour may vary. Drivers must learn how to respond quickly in such varying situations. To do this, drivers may internalize different estimates of precision (i.e., rates of variability) for different classes of agents (e.g., children might be more likely to cross the street without warning), and in turn, when travelling, drivers will re-adjust their expectations in light of local cultural variations in official rule-obeying (e.g., in a country where people are more likely to jaywalk). In addition to the internal updating of precision estimates, one can think of epistemic affordances as encoded in the social-ecological niche (Constant, Ramstead, et al. 2018b), in the patterned cultural practices that direct the epistemic foraging of agents (Ramstead, Veissière, and Kirmayer 2016), and in the specifically constructed aspects of the material environment (Constant, Bervoets, et al. 2018). For instance, drivers and pedestrians learn not only how to assess the information afforded by traffic lights, but also how to leverage the traffic light's probable influence on others to improve the quality of their assessment (Constant, Bervoets, et al. 2018), e.g., checking that the bus driver can see his red light, before stepping out onto a pedestrian crossing.

Responding to a culturally constructed niche depends on a developmental history of learning to negotiate similar niches (a developmental history that is shared with all conspecifics within the same econiche). In the process of development, however, humans not only respond to niches but take part actively in their (re)construction. For example, based on the frequency of traffic accidents at an intersection, the location or timing of traffic lights may be modified by collective action. This (re)construction of the niche occurs in more rudimentary ways constantly throughout the development of individuals and groups in local niches.

From the point of view of the FEP, *developmental niche construction* can be viewed as the process whereby agents make their niche conform to their expectations (Constant, Bervoets, et al. 2018). Developmental niches are the set of exogenetic, physically and behaviourally-grounded resources necessary to guide the reproduction of the adaptive life cycle (Stotz and Griffiths 2017; Stotz 2017). Because actions are guided by salience, and change the physical architecture (and epistemic affordance) of the environment, they tend to make the niche a good statistical 'mirror' of the agent's epistemic foraging, functional anatomy, and, ultimately, brain-based expectations (Constant, Ramstead, et al. 2018) (Figure 3). In short, if we all act successfully to minimise uncertainty our econiche will become inherently more predictable – if, and only if, epistemic affordances become encultured.

The exploitation of regimes of attention – encoded in the niche – is especially useful to track regularities unfolding over longer time scales of the history of a community, whose variability would be harder to assess over the timescale of an individual's perceptual and procedural learning. In humans, the epistemic affordance offered by niches constitute *epistemic resources* that shape learning, and shared cultural practices (D. D. Hutto 2012; Roepstorff, Niewöhner, and Beck 2010), as well as social relationships necessary for cooperative activities like breeding of animals (Burkart, Hrdy, and Van Schaik 2009). Many of these epistemic resources involve specific kinds of patterned cultural practice that we associate with regimes

of attention (Hutto 2012; Roepstorff, Niewöhner, and Beck 2010; Burkart, Hrdy, and Van Schaik 2009; Veissière 2016). These epistemic resources are states of the environment that, when repeatedly engaged by agents, shape their neurally encoded precision and salience expectations, and thereby, direct their future patterns of attention, epistemic foraging and learning, and subsequent patterns of engagement through perception and action. Epistemic resources help agents learn (from others) how to attend to or forage the niche for relevant affordances, and how to weigh the cues associated with different affordances. Epistemic resources allow the agent to track and evaluate the relevance of more abstract, temporally extended, stable, and general statistical regularities structuring agent-niche relationships, like conventionalized patterns of interaction shared among multiple agents.

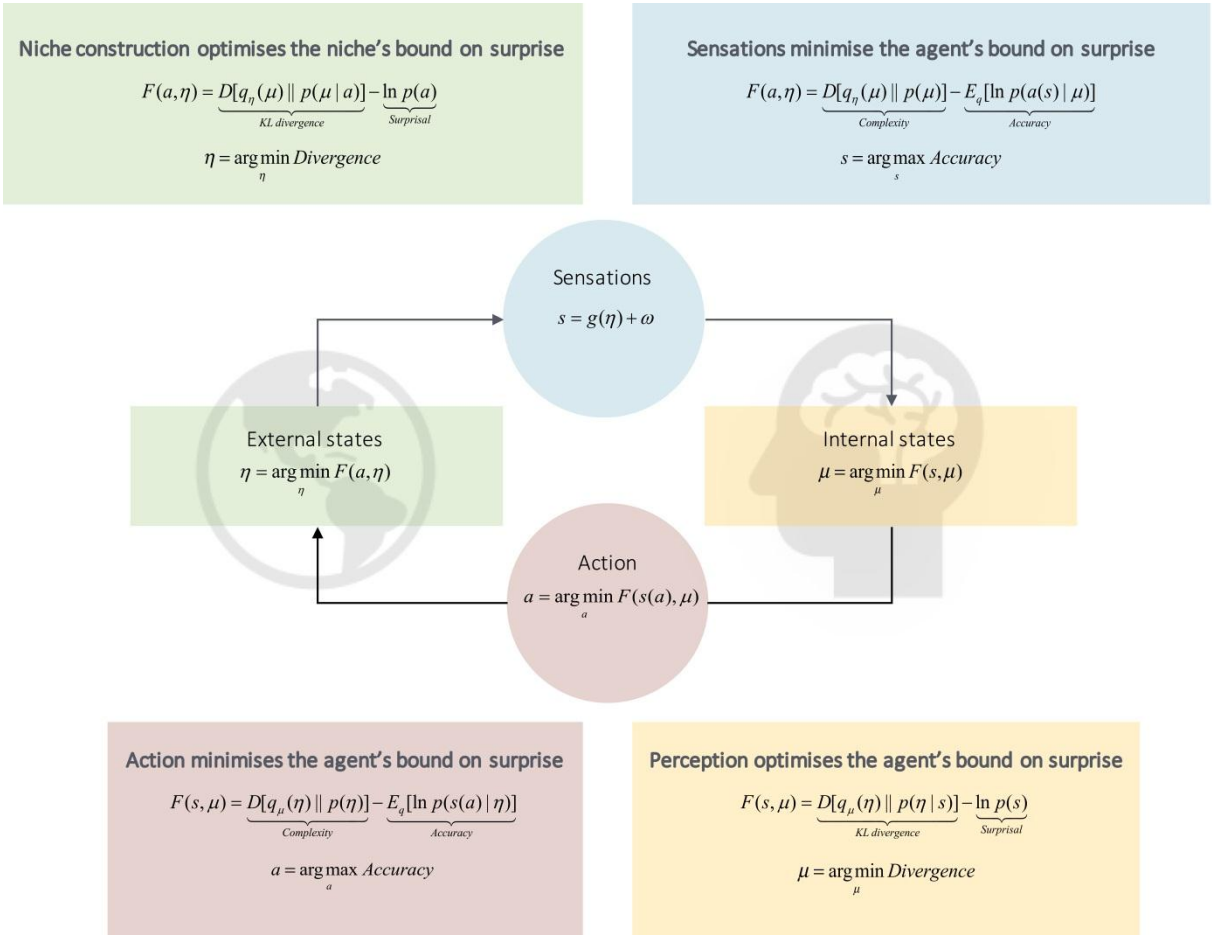


Fig. 3. Summary of the Variational Approach to Niche Construction. As in Figure 1, internal states and action change to minimise free-energy based on sensations and beliefs. Heuristically, one can think of niche construction as the process whereby the agent's action creates a symmetry between internal and external states. The agent changes the statistical structure of the world as it acts on the world. The statistical structure of the world here simply refers to the actual probability of finding some causes of outcomes at a given location in the environment (e.g., the bread being the cause of pleasant smell in the bakery). From the point of view of niche construction, such probability changes as a function of the agent's action, and in a way that is consistent with the agent's beliefs. Indeed, a simple consequence of agents acting to optimise action based on beliefs is that the traces produced by agents' action will tend to be consistent with their beliefs. Another intriguing consequence of this is that, over time, traces in the world will effectively 'learn' agents' beliefs, in the sense that those traces will encode statistical regularities that relate to those beliefs. For instance, consider a well-worn path cut through the grass in the park. Such a 'desire path' encodes a robust probability that the location of the path in the environment will map onto the probability outcome 'being walked on'. The value of that probability mapping increases over time as people wear down the path. This means that changes in the niche mirror changes in agents beliefs enacted via action. With the mathematical apparatus of the free energy principle, one can model 'environmental learning' about the agents' action in the same way that one models 'agent's learning' of the environment's sensory causes. The only twist is that the quantities are inversed (compare blue and green vs yellow and red boxes). From the point of view of the environment's generative process, actions play the same role as sensations in the agent's generative model (see Constant, Ramstead, et al. 2018; Bruineberg et al. 2018) for a detailed mathematical description).

3.5. Learning cultural affordances under the free energy principle

Epistemic affordances are encoded by – or installed in – the environment, as repeated physical actions leave traces that change the structure of the developmental niche in ways that influence agents' expectations (e.g., "I can trust that by taking this trail, which other people have also taken, I will end up at the other side of the park"). Over time, these traces of the actions of other people (e.g., traffic signals, dirt paths across a park, shelters for hikers along a mountain trail) make certain affordances stand out as especially relevant. These are the affordances that yield highly reliable actions (i.e., uncertainty minimizing action, or actions that are expected to guide the agent towards goals or expected states) (see Figure 4).

In many situations, affordances based on the history of human action will be more salient that those that reflect simple optimization (e.g., cutting across a lawn might afford getting to the other side faster, but many people will walk along a winding path, even in the absence of other humans). The well-worn path reflects an implicit consensus among many previous walkers. Individualized expectations guiding behaviour in context may thus be inferred from a *continuum of expectations about other agents*, ranging from reflective to fully intuitive, and in turn, from actually present to probable and generalized others. Under the FEP, the dynamics and acquisition of all these expectations by groups agents are mediated by the very same inference mechanisms.

Developmental niche construction can be cast as an interactional process between agents and a shared environment, producing affordances that support the reproduction of a normative life trajectory, through the norm-guided development of each new member of the community (cf. Fulda 2017; Constant, Ramstead, et al. 2018b). These norms are implicit in the structure of cultural affordances in the specific local niches occupied by individuals at a particular

developmental age or stage. Individuals become attuned to the niches they discover or are directed to by others according to their age, gender, and other dimensions of social status.

These niches afford individuals epistemic resources for acquiring specific types of knowledge, skills, or dispositions to respond. In effect, the function of external mechanisms for evaluating epistemic affordances is to enable the emergence and *stabilization of epistemic resources*. The notion of epistemic resources relates directly to work on how cultural knowledge held by others in the community can reach into the hierarchy of processing at higher levels through linguistic or symbolic communication to install priors directly (Bengio 2014).

Epistemic resources, which underwrite epistemic affordance (either overtly through action selection or covertly through attentional selection; i.e., mental action), are stabilized through niche construction, in the sense that the niche comes to encode the expectations that enable the interaction with those affordances. Epistemic resources act as developmental anchors. In human social contexts, epistemic resources can be viewed as shared expectations and cultural affordances that become available to a group of agents, as expectations that 'sediment' in public places, practices, and affordances that are repetitively and reiteratively engaged by groups of agents. This process involves feedback or looping effects and hence is self-reinforcing over time. For example, the grass patch on a street corner solicits cutting across, and over time and in turn, as it is worn down by many walkers, comes to afford a "desire path" (Ingold 2016).

One might ask whether the story should not be told the other way around. It might be that dirt trails allow for cutting across the park, but only later, solicits a 'desire path', as it is only once the agent has acquired the cultural knowledge that the path can be traversed that it can become 'desired' as something that the agent wants to engage. Precisely what is at stake here

is the virtuous circularity and bootstrapping operative in social learning – which must go from simple to more complicated. On a phenomenological level, what is being challenged is that the world calls to us in specific ways prior to the desires installed by culture – in cutting across the path, the unstated background of desire might have to do with getting somewhere we want to be more quickly, with enjoying transgressing the rule of walking (only) on sidewalks, or simply the aesthetics of walking along a dirt path. Hence, it is not self-evident that one can consider a desire path or for that matter, any cultural object, as a cultural affordance until some way of engaging the world has been acquired.

Affordances have been proposed to explain how skilled agents manage to engage their environment without having to know how their environment 'works'; i.e., to employ learnt representations, or to acquire representational contents. The variational approach furthers this line of thinking by distinguishing mathematically action that is selected by the agent and the *affordance* of action for the agent. In effect, the FEP allows us to formulate a *principle of most affordance*; that is a version of the principle of least action from physics, applied to the adaptive behaviour of groups of organisms living together in a niche (Ramstead et al. 2018). The action with the most affordance, the one that solicits the organism most (i.e., the one associated to the least *expected free energy*), is the one that ends up selected by the organism.

The cultural affordances framework suggests that acquiring the ability to leverage conventionalized affordances means acquiring a regime of attention. The regime of attention is not some specific content that one learns, but a mode of attending to and actively sampling the world, through a generative process that involves (overt) motor behaviour and the (covert) tuning of neural gating via expectations about precision, as well as culturally patterned search strategies for salient information, which are 'shared' to some extent by all individuals of a local culture.

The idea behind the desire path as a cultural affordance relies on and extends the notion of regime of attention by highlighting that *epistemic affordances* depend not only the brain, but also on features of the environment (see Figure 2)[6]. The desire path, as a cultural affordance, enables skilful pre-reflective engagement. This can often happen without the agent having to know the content of the specific artefact from the start. For instance, I might be late for my train, and following that trajectory through the park might be a good solution to catch my train on time. In that scenario, there is probably very little *content* involved with about where *exactly* the path will lead. Rather, there is (i) *an expectation* on the part of the agent, (ii) a *solicitation* on the part of the environment, and between those, (iii) an embodied *history* of agent-niche interactions (i.e., the traces left by repeated actions), which increases the likelihood of the path leading to a commonly experienced goal (e.g., the other side of the park). This history of cycles of expectation, solicitation, and action, encoded in cultural affordances, supports individuals' intuitive, culturally meaningful response to environmental cues. Under the TTOM model, when individual agents do not know quite what is situationally appropriate, their behaviour switches to *epistemic foraging*, in which agents will preferentially sample whatever other, relevant agents sample as well.

A large part of the social learning enabled by the developmental niche is mediated by shared attention (Tomasello 2014). For example, once a path is worn in the grass, implicit shared attention and expectations that others also intended to do so will prompt followers to walk along the path. This will hold even for paths that are not otherwise efficient, even if a less costly path is available – and, in some instances, this holds even for paths with uncertain trajectories or end-points. Of course, most of the traces of human activity are not paths on grass, but the affordances provided by institutions, archives, and repositories of knowledge, plans, and protocols. Regimes of attention provide ways to locate, attend to, and engage these affordances in a wide variety of structured cooperative activities (Malafouris 2015).

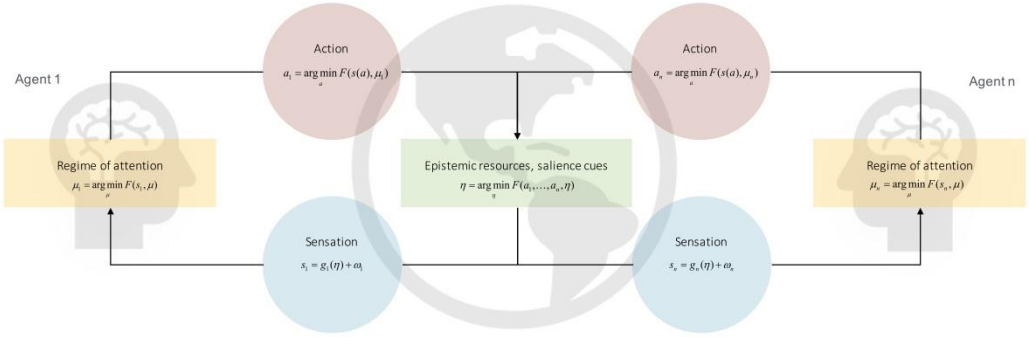


Fig. 4. Thinking through Other Minds (see Figures 1 and 3 for the equations). This figure depicts the loop between action, sensations, and niche construction that lead to the acquisition and production of cultural habits, and to the inference and learning about other minds. The shared epistemic resources in the constructed niche (i.e., external states modified by actions from agents 1 to n) and the regimes of attention (i.e., internal state) constitute the domains of statistical regularities that tune to one another via the physical engagement of the niche. Those domains are finessed (i.e., mutual learning of internal and external states) by a community of practices (agents from 1 to n) over ontogenetic (e.g., over development) and phylogenetic time scales (e.g., via the inheritance of material resources). The learning and deployment of internal and external domains of statistical regularities is what we call 'Thinking through Other Minds' (TTOM). TTOM entails, and depend on, the production of culturally patterned practices. Cultural practices and associated artefacts are epistemic resources that guide the attention (and learning) of members in the community by shaping sensory perception.

3.6. Why human thinking is always already thinking through other minds

Homo sapiens evolved to rely on bodies of accumulated cultural knowledge and skills for survival (Tomasello 2014; Sterelny 2012; Henrich 2015). We shape each other's learning through specifically adapted cultural practices (regimes of attention) that allow individuals to enact recursively nested forms of intentionality. This includes the capacity to view ourselves

through the eyes of another in a kind of reciprocal aboutness: e.g. “What would Mother expect me to do?”. After childhood, typically, these ways of thinking about oneself are internalized, encoded and expressed as “What should I do?” or “What am I expected to do?”. Recent research on mind-wandering suggests that most of our spontaneous mental life is dedicated to rehearsing social scenarios (Poerio and Smallwood 2016). In their recent ‘interactionist’ account of the evolution of human reasoning, Sperber and Mercier (Mercier and Sperber 2017a) review a wealth of experimental evidence to support the claim that humans best solve problems and optimize individual intelligence collectively in dialogical and argumentative contexts, which may extend to hypothetical, ‘silent’ scenarios. While no large-scale evidence is available on what so-called ‘silent reasoning’ entails in individual human heads, Sperber and Mercier conjecture that most silent reflective ideas are generated through the rehearsal of arguments with, and justifications to others. Even solitary thinking, on this view, is a rehearsal for bona fide social interactions with peers.

Recent work in the philosophy of psychiatry also supports the hypothesis that solitary human cognition is social through and through**.** In their cultural and evolutionary account of the origins of psychosis, for example, Gold and Gold (Gold and Gold 2015) propose that the many kinds of delusions described in the literature on psychopathology (i.e., persecutory, grandiose, erotomanic, control, thought, somatic, nihilistic, reference, guilt, and misidentification) share one broad, overarching theme: a concern with one’s relationship to *other people*. Hence, all known delusions can be recast as statistically improbable interpretations of, and expectations about, one’s experiences *in relation to others*.

For a species such as *Homo sapiens* that evolved to rely upon cooperative and highly elaborate coordinated action, expectations about folk psychology (or probabilistic inferences about the way other people think and reason and what they expect of the world) are at least as important as, if not more important than, expectations about statistical regularities that

characterize the physical world. In other words, in a world populated by creatures 'like me', most of my expectations call on the prior belief that 'I am like you and you are like me – and you believe that I am like you and you are like me' and so on. In effect, the world of human experience is always already mediated by, and filtered through, the 'lens' of expectations about another's expectations.

The expectations that *Homo sapiens* have leveraged most over their phylogenetic history involve the capacity to '*outsource*' *cognition to relevant others* (people, artefacts, practices, and institutions). In other words, human beings outsource to other humans many of the evaluations of salience that they employ in their engagement with their worlds, which allows others to perform culturally relevant tasks (Tomasello 2014). Indeed, it is precisely these evaluations by others that make worlds 'meaningful' for humans. To exploit this cooperative cognitive task sharing, humans agents explicitly and implicitly bestow trust and assign authority to others—both individuals and institutions—acquiescing to and leveraging cues (physical, culturally meaningful *signs*) associated with reliability, authority, prestige (Henrich 2015).

What distinguishes between different human phenotypes is the priors under which they are operating, and which guide adaptive behaviour. If we consider the dynamics of human TOM abilities in this light, the process of TTOM consists in inferring the priors or expectations that guide the beliefs of another agent or group of agents. Provided that agents can solve the inference problem about the sort of person that their interlocutors are, and provided that they have a model of their conspecifics' prior beliefs, then any one agent can leverage their own action (policy) selection mechanisms under the prior beliefs of their fellows to infer the mental states of their fellows (and, indeed, their own mental states).

Epistemics get into the game when this inference is made more difficult by a lack of shared priors. Thus, the cues that emerge from niche construction can be nonspecific cues that tell agents about what is situationally appropriate to do (but which could be done in a solitary way, like stopping at a red traffic light), or very particular cues that provide information about the priors of other agents -- which coincides with mindreading and properly thinking through *other minds* (e.g., I have a prior about you having a prior about me stopping at the red light, and crossing at the green light – and, hence, that you won't run me over). The process of inference is made easier by the availability of cues (that shape regimes of attention) that tell agents 'where to look'; i.e., that allow one to leverage where others are looking to determine where oneself should look. For instance, if I don't know when to cross at the intersection because I am not familiar with the colours used by the traffic light system, I can guide my action by relying on epistemic cues that have been shaped by (presumably adaptive) cultural practices such as the ways people around me act in context (e.g., other agents' behaviour or gaze patterns).

The TTOM model accounts for the ways in which human agents outsource their policy selection to relevant others and to aspects of their material niche. In this sense, our model covers cases of cultural cognition that range from the lone encultured agent acting in conformity with the cultural norms that they have internalized – which involves inferences only indirectly about and through other minds – to full-blown cultural engagement with other human agents, that requires (implicit and explicit) inferences about the minds of other humans. Given the nature of their inferential systems and the way they learn generative models according to TTOM, inferences about my own generative model can be leveraged, and in effect, is always being leveraged, to make inferences *about others like me*. Inference about one's own mind is always mediated and made possible by inferences about the minds of others.

4. Addressing TOM critiques with TTOM

According to TTOM, human agents organize most of their behaviour as a function of *what they can infer from other human minds*. Humans find guides for action by picking up on statistical regularities in the realm of *folk psychology*, which identifies the most relevant states of the external world, as well as the most relevant sources of inferences about the shared social world. Our framework recognizes the contribution of the varied approaches to human TOM abilities outlined in the first section and offers a compromise position.

4.1. Response to the cross-cultural critique: TTOM is universal for *Homo sapiens*, but realized through cultural niches

We agree that folk notions of personhood vary across culture, and likely exercise specific constraints on automatic perception and social coordination through *normative* social learning (e.g., McGeer 2007). While folk notions of the locus of personhood and agency vary broadly between groups and historical periods (e.g., to include a soul, brain-mind, heart-mind, or external agencies like gods, ancestors or spirits), we question the extent to which communication and coordination would be possible without a species-wide intuitive notion of *propositional psychological interiority* (which may be postulated and enriched in different ways culturally).

The example of 'silent thinking' during courtship, reported from ethnographers of the Korowai (Stasch 2009), is telling. In everyday human experience, affectively charged situations such as "I wonder if she really likes me" abound, and likely emerge in infancy without recourse to language or explicit mentalizing, as humans form mental models of other agents in their life. Indeed, developmental psychologists have shown that 15-months-old infants are able to take into account the false beliefs of other agents (Onishi and Baillargeon 2005) and that the ability to attribute goals to any entity (living or not) that appears to be

animate emerges as early as 5-months (Luo and Baillargeon 2005); see (Mahajan and Woodward 2009) for different results).

Additional cross-cultural and developmental findings support the view that intuitive dualism (Jack 2014) (or the folk tendency to situate personhood in an intangible psychological interior) is likely a cross-cultural universal that does not require specific cultural immersion in Cartesian cultures (Chudek et al. 2013). As Paul Bloom has argued (Bloom 2005), children across cultures can readily understand a story about a prince becoming a frog without explicit enculturation into folk Cartesianism.

As we argue below, TTOM makes no ontological claims about mind-body dualism; we simply point out from experimental and ethnographic evidence that coordinated action in human sociality *does* rest on the universal human cognitive capacity to understand others as having goals, beliefs, desires, and intentions that may be different from their stated ones (what we call 'propositional psychological interiority'). At the core of this cognitive capacity is the process of active inference mediated by processes of developmental and selective niche construction, which in humans, scaffold complex sets of prior beliefs encoded in sites across the brain-body-environment-others system. Hence 'mindreading' sometimes requires explicit deliberation (something resembling "Theory Theory") and can at other times can be automatically intuited through simulation (in forms of embodied and extended cognition).

4.2. Response to the embodiment critique: TTOM is grounded in the bodies of self and others

Anxieties around dualism in current cognitive science reflect a common confusion between *normative* and *descriptive* commitments on the part of philosophers and cognitive scientists. Although dualism as a scientific description of the relation between the mind and body is mistaken, it does not follow that our theorizing about other minds should not consider folk

dualist thinking as a normative and very real phenomenon that shapes every day and scientific thinking. As an illustration, even psychiatrists who espouse an integrative, monistic view of mind and body employ a naive dualism in assessing vignettes of problematic behaviour as indicating either deliberate action (rooted in individual psychology, and hence, blameworthy) or as accidental, due to malfunctioning biology of the brain (Miresco and Kirmayer 2006) – as though these two causes were grounded in distinct mental and bodily processes. Our best theories about folk social cognition ought to reflect that dualism, on pain of descriptive inadequacy.

TTOM, to be sure, does not make ontological claims about the nature of mind as separate from the body. We simply offer that, as a matter of universal human epistemology, patterned cultural practice involves an ability to make inferences from, through, and about other minds, as propositional processes — indeed as inferential processes. In some cases, folk theorising about dualism may simply be a useful tool to both generate and inquire on such practices (e.g., through dialogues in clinical setting). TTOM formalises the inferential structure of such folk theorising.

The ability to infer each other's expectations, which makes human cognition, sociality, and culture possible at all, ranges from the fully explicit to the fully automatic depending on the situation. In our model, this ability depends on the learning of a spectrum of expectations encoded across the brains-bodies-environment-others system that underwrites regimes of attentions. The FEP is unique here in its ability to account for *inference and dynamics* as two sides of the same coin, and this is what allows TTOM to overcome the sharp dichotomy between internalist and externalist approaches to TOM abilities. Under the FEP, all systems dynamics are inferential, and inference is itself dynamics; namely, the dynamics of sentient

systems are a gradient flow over free energy (Friston 2010; Ramstead, Badcock, and Friston 2017). Since free energy is a measure of the complementarity between the organism and the niche, in terms of a generative model of the relation between them, any dynamics formulated in terms of the FEP are ipso facto *inferential dynamics* that pertain to the *self-organization of information flows in sentient systems*.

Rather than describing cultural differences in the folk models (including Western philosophical models!) of social cognition in 'either/or' terms (either dualistic or not; focusing on explicit intentions or focusing on resonance in action), we propose to situate these differences on a continuum of *hypo-cognition* to *hyper-cognition* of intentions; see (Duranti 2015). The notion of hyper- and hypo-cognition has been explored in the context of cultural variations in emotions (Lévy 1984; Levy 1975). The degree or depth of cognitive elaboration of emotion serves individual and social regulatory functions. As a matter of normative concern, cultures vary in the kinds of emotions people are encouraged to cultivate or suppress, thereby allocating attention, attributing meaning, and patterning behaviour in ways that constitute specific codes of conduct or expression, modes of experience, and folk explanations that account for behaviour.

4.3. Response to the cooperativity critique: TTOM is built on the developmental scaffolding of cooperativity

Shedding light on a cross-cultural continuum of normative commitments to the hyper- and hypo- cognition of intentions may also help resolve the Machiavellian-mutualist debate on the evolution of human cognition. It seems self-evident from the human record that our species is capable of both selfishness and altruism as a matter of individual, situational, and cultural variation - but also that the scaffolding of 'altruism' proper clearly follows an evolutionary and developmental trajectory. Tomasello (Tomasello 2009), for example, proposed the Early

Spelke, Later Dweck Hypothesis[7] to describe children's gradual immersion into social norms that harness and enhance their natural capacity for adjusting their behaviour to what others expect of them.

Rather than start from a specific commitment to one normative position (e.g., "humans ought to be altruistic"; "humans ought to act in rational self-interest"), our account recognizes these varied possibilities inherent in human behaviour, and stresses the importance of specific cultural practices in patterning behaviour to elaborate either side of the selfish-altruistic continuum.

Hrdy herself, as a proponent of the mutualist argument, has stressed the importance of developmental environments, such as collective parenting, in providing rich (or impoverished) opportunities to form bonds and learn to relate with multiple attachment figures — a process she describes as crucial in the development of social cognition, emotional regulation, and empathy (Hrdy 2011). In Hrdy's account, our 'proximity' to the kind of selfish intelligence found among chimpanzees is a matter of ontogenetic contingencies at least as much as evolutionary 'distance'. Indeed, the capacity to engage in nuanced, compassionate, other-regarding action is increasingly understood to be dependent on language, explicit teaching, effortful deliberation and practices, and to be distinct from (though perhaps developmentally scaffolded on) the innate capacity to imitate and follow others and favour one's narrow in-group (Bloom 2017).

Contemplative practices of loving kindness meditation, for example, entail the explicit enrichment and effortful rehearsal of one's mental models of others, which eventually become automatic through practice (Lutz et al. 2008; Lebois et al. 2018). The linguistic (narrative) elaboration of these models may be essential to their extension to include members of out-groups, the whole of humanity, or even to all sentient beings. These varied examples point to

the importance of both implicit and explicit mentalizing mechanisms in the mediation of human cognition and cultural practice.

TTOM supports current mutualist, cultural intelligence, or 'dual-inheritance' accounts that emphasize the co-evolution of human cognition and culture (Henrich 2015; Tomasello 2014). Rather than to discount Machiavellian and other 'selfish' accounts of these processes altogether, we suggest that what one might call *extended mutualism* – i.e., large-scale cooperation – and the ability to leverage a large repertoire of shared expectations to guide group action – arises because of the match between naturally and culturally selected dispositions to acquire cultural abilities (e.g., mindreading abilities) and inherited developmental conditions enabling the (re)acquisition of these abilities. Selected, or evolutionarily old dispositions constitute a cultural learning 'start-up kit' of sorts (Heyes and Frith 2014; Heyes, 2018), which includes the kind of neural machinery that underwrites attention and the estimation of salience, leading to the acquisition of shared expectations (see Figure 2).

At the developmental time scale, inherited cultural practices enable the learning of shared expectations via the patterning of those evolutionarily old dispositions. This emerges via agents' engagement with epistemic cues that undergo processes of cultural evolution through developmental niche construction activities, which filter what persists across generations as a function of the success of the behaviours they afford (Laland 2018) (see Figure 3).

This sets up a cycle of mutual fitting between individual and niche. For instance, in a circular fashion, I can trust the learning biases provided by my caregiver – and more specifically, the cues they provide through their gaze direction, pointing, gesturing, etc., towards salient situations. I am licensed to do this because patterns of offspring-caregiver interaction have been filtered and fine-tuned through gene-culture coevolutionary processes, and developed in

specific cultural norms, signs, places, and practices over historical time – all in the service of guiding the learning of salience; i.e., to guide the learning of what is adaptive in the local cultural context (e.g., "listen to and copy this high prestige individual because prestigious individuals are typically the ones that have succeeded in the past"). Put another way, one can trust learning biases since biases indicate action policies selected by other agents 'like me', these must have the most adaptive for creatures 'like me'.

On our account, cognition and culture are largely synonymous for humans, as both are predicated on the capacity for shared expectations. Priors leveraged and finessed through active inference, and the folk psychology they specify (i.e., what we expect others also to expect) constitute the central domain of statistical regularities that ground humans' models of their world. This domain of statistical regularities that we call TTOM specifies the mechanistic processes that drives the implicit acquisition of culture over development.

5. Concluding remarks: The future of TTOM

5.1. Future research

We have argued that the pervasive influence of culture, through widespread shared expectations, institutions and practices, can be cast as a process of co-constructing and responding to a shared set of affordances. Human engagement with cultural affordances is enabled by (often implicit, recursively nested) expectations about other relevant agents' expectations. These expectations are acquired by agents through immersive participation in the practices that define their shared way of life, in a process which gradually takes hold in ontogeny through regimes of attention and niche construction.

The human mind is optimized for outsourcing information to other human minds in order to function in a niche that requires the shared, coordinated pursuit of joint goals. Error and

surprise minimization in large-scale social systems hold because *individual* human minds are coupled to one another in an environment of other minds. This kind of 'extended mind' is distinctive to human beings due to the capacities for culture (i.e., regimes of attention, linguistic communication and installation of higher-order priors, multiscale cooperation, declarative memory/historicity, collective norms and goal setting) that are made possible by human nervous systems (Clark 2008; Menary 2010; Clark and Chalmers 1998; Sutton 2010).

If we have been successful in presenting our account, however, from an FEP point of view, it should also be clear that humans think, feel, imagine, and act in ways that are only possible because they are afforded by the niches they inhabit and co-construct, and the cultural practices that make up their shared form of life, and which all serve to enculture human agents (Ramstead, Viessière, and Kirmayer 2016; Constant, Bervoets et al 2018; Constant, Ramstead et al. 2018; Ramstead, Constant et al 2018). Even the collaborative construction of new niches, which allows the exploration of new modes of experience and the improvisation of new forms of cooperative action, depends on the cultural scaffolding of a relatively stable set of shared expectations and regimes of attention through the cognitive tools or gadgets of narrative and metaphor (Lakoff and Johnson 1980; Heyes 2018) and the social organization that constitutes particular niches or communities.

TTOM is a generic active inference (a.k.a. FEP, or variational) account of the acquisition of culture and mindreading abilities. We have designed TTOM as a guide for the production of *testable models* in related domains. While TTOM per se would be difficult to test (due to its generality), one can derive specific, integrative models from TTOM to study specific forms of socio-cultural dynamics. A good example of a testable model derived from TTOM is the theory of Regimes of Expectations as applied to the study of social conformity (Constant, Ramstead et al. 2019).

Social conformity refers to the deference to social norms such as embodied by other agents. From the point of view of social psychology, social conformity is one possible response to social influence of epistemic, trusted others (Asch 1956). From the point of view of cultural evolution, in turn, social conformity is viewed as an adaptive social learning strategy in uncertain environment (Morgan and Laland 2012).

The theory of Regimes of Expectations integrates the perspectives of social psychology and cultural evolutionary theory by modelling social conformity as a process that obtains through the intergenerational finessing of environmental cues that guide social learning over development. Social learning that is aided by these cues, in turn, allows the active inference agent to perform action selection in a fast and efficient way in uncertain contexts by *leveraging trusted others* (either through material cues that stand as culturally-signalled proxies for other, relevant or prestigious minds, or directly by copying such individuals). These trusted others are defined as 'deontic cues' (Constant et al. 2019).

'Deontic cues' in this model are context-specific epistemic resources (as defined by TTOM) that enforce an obligatory response to the context that embeds them (e.g., a red traffic light enforcing stopping behaviour). The theory of Regimes of Expectations models social conformity as an active inference process of action selection that operates via the estimation of the epistemic, pragmatic, and also *'deontic' value* of action; which is the type of value learnt through the engagement of deontic cues. The deontic value is essentially the value of an action policy specified by the shared beliefs and preferences of a sociocultural group.

In line with the sort of specific models that can be derived from TTOM, the theory of Regimes of Expectations as applied to the study of social conformity integrates externalist approaches (e.g., cultural evolutionary approach) and internalist ones (e.g., the social

psychology approach) by describing the cultural domain of statistical regularities optimised through active inference and governing action selection.

The theory of Regimes of Expectations as applied to the study of social conformity makes specific predictions that stems from the TTOM model, namely that: (i) social conformity leads to more efficient cognitive processing and policy selection (e.g., as conveyed by psychophysics measurements like reaction time) in the presence of deontic cues (epistemic resources in TTOM terms); (ii) conforming actions minimise variational free energy over time more efficiently in social context – since regimes of attention will be optimised for zeroing in on social information conveyed through deontic cues; (iii) deontic cues reproduce conformist biases in cross-cultural between-subjects designs, but fail in within-subjects designs – i.e., not all deontic cues will elicit social conformity for participants with culturally diverse background due to the influence of culture-specific regimes of attention.

5.2. Limitations

Because it is based on the FEP, TTOM provides a mathematical formalism that can be used to model the effects of cultural affordances on adaptation to specific kinds of social niches. The model needs to be further elaborated to deal explicitly with the many varieties of cultural learning and regimes of attention. These include the distinctively human functions of narrativity that entail the linguistic and symbolic hierarchical installation of higher-order priors (Bengio 2014). For instance, this will include culturally shared expectations about the cause of sensory observations (e.g., the prior belief that ‘the slap I received on my wrist was caused by my belief that it is permissible to reach for the cookie jar, which motivated my action, which then led to the slap, which indicated it was not’. In this sequence, the slap not only conveys a social norm but in itself reflects the broader social norm that it is permissible to intervene in childrearing in this fashion — these overarching norms are learned over time

within a particular niche and may change, for example, with migration to a new sociocultural context, with serious consequences for how one (mis)reads (culturally conventional or permissible) affordances). In modelling an active inference agent, such structures of high-order priors could capture the potential for reflexivity and self-reference that gives human cultural-linguistic cognition its unique reach (Taylor 2016).

The free energy minimising dynamics described above involve feedback processes that tune organismic expectancies to fit local environmental contexts and therein minimise surprise and uncertainty. Accounts of enculturation tend to suppose *stable social contexts*, and the FEP assumes a kind of optimization that depends on *stability in adaptive contexts*, but the reality (especially in the context of cultural interactions and contexts) is often one of *constant change*. Thus, realistic models of human cognition in context will require taking into account cultural mobility, hybridity, and the cognitive effects of the constantly changing social niches that reflect cultural co-evolution. Ultimately, models based on conservative processes like the FEP model need to address the significance of historicity and contingency in the emergence and evolution of cultural systems.

Among other potential domains of application, our model has implications for psychiatry. One interesting path towards experimental verification builds on recent proposals for a *computational psychiatry* (Montague et al. 2012; Friston et al. 2014; Adams, Huys, and Roiser 2016; Huys, Maia, and Frank 2016). In brief, computational psychiatry aims to leverage computational techniques in order to better phenotype various psychiatric conditions, such as psychosis (Adams, Huys, and Roiser 2016) and autism (Constant, Bervoets, et al. 2018). Characterizing individual and group variations in the capacity to leverage TTOM, and the ways in which human agents adapt to their ecological niche, could reveal an important set of dimensions for such diagnostic frameworks. One could, for instance, consider individuals who experience inference about the sort of person they and others are in a way markedly

different from the neurotypical population (e.g., people with autistic traits). One could recruit participants who score high and low on the autistic spectrum, and to test their relative ability to make inferences and predictions about others based on the ability to leverage information about gaze direction; or vary the context in which they deploy such inferences, to study the coupled dynamics between context and cognition that is typical to such individuals (Constant, Bervoets, et al. 2018).

Other conditions could be studied in this manner as well, shedding light both on TTOM as a general cognitive architecture and on these specific conditions. Higher rates of schizophrenia and psychosis among migrant populations might also be an excellent lens to approach such phenomena. Indeed, the careful study of such populations highlights the need for an interactional view of how sense of self and functioning may be destabilized by migration – to a new niche that has specific affordances for people of colour (Kirmayer and Gold 2011, 2012; Kirmayer, Lemelson, and Cummings 2015). Depression might also be a useful phenomenon to consider, as it is an interactional phenomenon that involves complex inferences about self and other that is aggravated by retreat from the social niche, now perceived as lacking positively valenced affordances and occupied by other minds with intentions that are hard to understand, and which may in turn aggravate the condition itself (Wang et al. 2008; Baldwin 1992). This kind of work could inform a formal phenotyping of psychopathology based on the TTOM model.

Finally, while arguing for the applicability of the FEP to the puzzle of the acquisition of cultural practices, knowledge and grammars, we caution against describing cultural ensembles as autonomous systems that maintain their organization and structural integrity through allostasis and homeostasis (Veissière 2017). Adaptation rests on an ongoing process of predicting events, engaging with the environment, and adjusting expectations in response to feedback from the world (including the body and other creatures). This occurs through

constant transactions with the environment and, in the case of human beings, that environment is fundamentally cultural and social – constructed with, and inhabited by, other people with whom the individual agent must cooperate if they are to survive. This cooperation is itself patterned by cultural knowledge, skills, norms, institutions, places, and practices that have their own history and contingency.

Box 1. Glossary of key terms

Active inference: Active inference is the process whereby organisms learn the statistical structure of their environment through the selective sampling of predicted or expected sensory information (aka, action), based on perceptual inferences about the cause of the sensory input (aka, perception). The process of active inference realises the free energy principle. In active inference, everything that can change, changes to minimise variational free energy, which is a statistical measure of the mismatch between organism and environment. This mandates actions that minimise expected free energy following an action; namely, actions that resolve uncertainty.

Affordance: Generally speaking, possibilities for engagement with an ecological niche that are defined in interactional terms, as a relation between features of organisms' environment and their own abilities.

Attentional salience: The degree to which uncertainty is reduced under a particular course of action. Mathematically, salience is known as expected Bayesian surprise, information gain, intrinsic motivation and epistemic value. Salience underwrites epistemic affordance.

Attentional selection: Calibration or weighting of the precision (inverse variance) of sensory evidence, or prior beliefs.

Conventional affordance: Affordances that agents can engage by skilfully leveraging explicit or implicit expectations, norms, conventions, and cooperative social practices.

Cultural affordance: The kind of affordance that characterise the human niche. Cultural affordances depend on shared expectations that are acquired over development (i.e., through enculturation and social learning). Cultural affordances come in two flavours, which form a spectrum from the more innately specified to the more learning-dependent: natural and conventional affordances.

Epistemic affordance: One of the two components of expected free energy that determines action selection. Epistemic affordance quantifies the extent to which a particular way of actively sampling the world reduces uncertainty about the state of the world or its statistical regularities.

Epistemic authority: A symbol, person, cue, or feature of the environment (usually associated with prestige, status, and group affiliation) that signals salient, high-quality, uncertainty-reducing information in a given cultural context, and as such possess the 'power' to guide attention, enhance credibility, and prescribe action (e.g., biomedicine and neuroscience possess high epistemic authority in current culture; "The Guardian" newspaper possesses high epistemic authority for liberals, as does "Fox News" for conservatives).

Epistemic foraging: The agent's uncertainty-resolving behaviour. Epistemic foraging disambiguates Bayesian beliefs about a situation in order to be better poised to exploit the pragmatic value of action (i.e., value that relates to the sensory preference of the agent).

Epistemic resources (a.k.a. cultural affordances): Cues that are encoded in external states of the ecological niche (e.g., material cues and other agents), which guide epistemic foraging and implicit learning of patterned cultural practices.

Expectations: Bayesian beliefs and preferences about external states of the world, which are operationalized as probability distributions.

Free energy principle (*FEP*): A principle of least action derived from information theory. The free energy principle states the minimal conditions that systems must meet if they are able to endure in a bounded set of states (i.e., if they are endowed with a phenotype).

Generative model: A probability distribution or mapping from beliefs about hidden causes to observed consequences; i.e., sensations. Technically, this is the joint probability of a sensory state and a (hidden) state of the world. Under the FEP, the generative model defines free energy gradients (a function of sensations and predictions under the generative model) and subsequent perception and action.

Natural affordance: Affordances that agents can engage by leveraging their innate phenotypical endowments.

Niche construction: The process whereby organisms (implicitly and explicitly) modify their ecological niches, such that the states of the environment come to encode relevant aspects of their prior beliefs, which they can leverage 'downstream' to optimise their adaptive behaviour and act in contextually appropriate ways. The 'Janus face' of active inference.

Pragmatic affordances: One of the two components of expected free energy in policy selection. Pragmatic affordance is essentially equivalent to expected utility in economics, and quantifies the extent to which an action policy conforms to the prior preferences of the agent (also known as pragmatic or instrumental value).

Regimes of attention: Patterned cultural practices whereby members of a group of people acquire and maintain shared expectations that *modulate* attention, *structures* salience, and thereby *guide* action (Figure 2); as well as the internalised patterns of attention that result

from the repeated engagement with such practices (e.g., as a group-specific affordance, it takes a regime of attention for the colour white to signify mourning for Hindus; it also takes a species-wide regime of attention for humans to feel invited by a path in the woods that signals the trace of other human's intentions).

Salience: Expected information gain under a given action.

Surprise: aka surprisal or self-information in information theory. This is simply the negative log probability of some state or event.

Thinking through Other Minds (*TTOM*): The *domain* of beliefs about statistical regularities (i.e., Bayesian prior beliefs) that are exploited in learning cultural affordances. This domain is primarily situated in the realm of *expectations that humans learn to form about other people in the niche*, that is, in the realm of *folk psychology*. TTOM is also the *process* of engaging others' expectations and inferences by leveraging this domain.

Box 2. The formal structure of the FEP model adds significantly to the general approach we outline in this paper in two ways.

1. *Conceptually*, the FEP provides us with an explanation *from first principles* of the processes involved in, and the adaptive value of, implicit cultural learning and mindreading abilities. It gives us a formal grip on the underlying dynamics of these two phenomena (for a schematic overview, see Figs. 1, 2, 3, 4 and the mathematical appendix). The main challenge confronting TTOM is that of making sense of the dynamics involved when agents' learn domains of socially relevant expectations – that are involved in the acquisition of culture – and how these domains are scaffolded from joint intentionality, basic perspective-taking abilities, and evolved attentional dispositions for learning from and through others. These domains are internal (e.g., neural scale) and external (environmental scale) to individual

agents. Without a formal apparatus, it is difficult to make sense of these multiscale learning dynamics or to examine how they interact. We employ the FEP to formulate TTOM for the simple reason that it is, to our knowledge, the *only* theory that has produced formal models (supported by computer simulations) of many of the cognitive mechanisms involved in the learning dynamics of TTOM, including, for example, action, perception, learning and attention (Friston, FitzGerald et al., 2016), visual foraging (Mirza, Adams et al. 2016), communication (Friston and Frith, 2015b), decision making, (Friston, Schwartenbeck et al., 2014), planning and navigation, (Kaplan & Friston, 2018), emotions (Joffily & Coricelli, 2013), curiosity and insights (Friston, Lin et al., 2017), and niche construction (Constant, Ramstead et al., 2018; Bruineberg, Rietveld et al. 2018).

2. *Empirically*, the FEP offers a set of equations that can be used to develop computational models of data acquired in studies of social interaction, in which implicit cultural learning and mind reading are at play. These models can then be used to identify new dynamics and make predictions which can, in turn, be tested in real-world situations. The scope of the current argument is limited to discussing the theoretical relevance of the FEP. That said, we can indicate candidate tasks to produce data amenable to FEP modelling. Notably, the different variants of two-person psychophysiology in social interaction studies (e.g., Timmerman, Bert et al 2012; Schilbach 2016; Lühe, von der et al 2016; Bolis and Schilbach 2017; Bolis, Balsters et al. 2017) are target modelling candidates, as they already rely on core principles of active inference and involve the manipulation of what we call "epistemic resources."

Endnotes:

[1]There are many ways of interpreting this haiku by the modern poet Mayuzumi Madoka. The shift in gaze might be seen as an experience of erotic presence or represent an awakening to sexism and self-estrangement. It also recalls a culture-specific experience of the self as a

performance (echoing the Japanese sense of always being on a stage (Heine et al. 2008)). At its core, though, the poem powerfully illustrates the fundamentally human affective process of seeing and feeling oneself through the perspectives (and desires) of another.

[2]Technically, an expectation corresponds to the average of a probabilistic belief or probability distribution. When the distribution is over (discrete) states of affairs, the expectation corresponds to the likelihood that any given state of affairs is true. Throughout, we will use beliefs in the sense of Bayesian belief updating or belief propagation which could be either propositional or subpersonal in nature.

[3]i.e., the *act* of deploying precision weighting to select sources of sensory evidence, often discussed in terms of *mental action*.

[4]The FEP is a variational principle of least action, like those that describe other systems with conserved quantities, e.g., in the Lagrangian formulation of Newtonian mechanics, in which energy and momentum are conserved (Coopersmith 2017).

[5]Intrinsic motivation is commonly used in developmental robotics to describe the epistemic value that reduces uncertainty (i.e., promotes information gain). In active inference, salience scores the reduction in uncertainty about transient states of the world, while novelty scores the reduction in uncertainty about the more stable parameters of a generative model. In short, *salience* is to *inference* as *novelty* is to *learning*.

[6]The epistemic, uncertainty-reducing aspect of this formulation comes to the fore when human agents need to figure what to do, more so than when agents are simply acting in accordance with the regimes of attention that they have internalised through enculturation.

[7]With reference to the works of psychologists, Elizabeth Spelke (who documents infant ‘core

knowledge' in the domains of intuitive physics, intuitive biology, and intuitive psychology) and Carol Dweck (Dweck 2013; Johnson, Dweck, and Chen 2007), who emphasizes the role of learning, experience, and rewards from adherence to social norms) (Olson and Spelke 2008; Spelke and Kinzler 2007; Kinzler, Dupoux, and Spelke 2007).

Appendix

This appendix describes the free energy principle in terms of a Bayesian mechanics that emerges from the existence of a Markov blanket in a random dynamical system at non-equilibrium steady-state. A Markov blanket is a four-way partition of states that define a self-organising system and its environment (i.e., a system that has self-organised to nonequilibrium steady-state). This partition comprises *internal* and *external* states $\{\mu, \eta\}$ that are separated by blanket states $b = \{s, a\}$. In turn, blanket states are divided into *sensory* and *active* states. In brief, the Markov blanket allows us to talk about internal states *representing* external states in a probabilistic sense. Heuristically, this means that one can ascribe probabilistic beliefs to internal states, in the sense that they are about something; namely, external states. This interpretation rests upon a *variational density* over external states that is parameterised by internal states:

$$\begin{aligned} \boldsymbol{\mu}(b) &\triangleq \arg\max_{\mu} p(\mu \mid b) \\ q_{\boldsymbol{\mu}}(\eta) &= p(\eta \mid b) \end{aligned} \tag{1.1}$$

This variational density arises in virtue of the blanket as follows: if we condition internal and external states on the blanket, then there must exist a most likely internal state for every blanket state. This means that there must be a conditional density over external states conditioned on that blanket state. At nonequilibrium steady-state, the flow of internal and active states can be expressed as a gradient flow on the *same quantity*; namely, the surprisal

(i.e., negative log likelihood) of states that comprise the system (Friston, 2013). We will refer to internal and active states $\alpha = \{a, \mu\}$ as *autonomous* because they are not influenced by external states:

$$\begin{aligned} f_\alpha(s,\alpha) &= (Q_{\alpha\alpha} - \Gamma_{\alpha\alpha})\nabla_\alpha \Im(s,\alpha) \\ \Im(s,\alpha) &= -\ln p(s,\alpha) \end{aligned} \tag{1.2}$$

These two aspects of a Markov blanket underwrite a Bayesian mechanics, in which we can talk about internal states holding Bayesian beliefs about external states – and autonomous states acting on external states, under those beliefs. We will first look at the underlying formalism in terms of a free energy lemma and its path integral form that speak to (i) the most likely flow of internal states (i.e., perception) and (ii) the trajectory of active states (i.e., action).

Lemma (variational free energy): *given a variational density:* $q_\mu(\eta) = p(\eta \mid b)$, *the most likely path of autonomous states, given sensory states, can be expressed as a gradient flow on a free energy functional of systemic states:* $\pi = \{b, \mu\} = \{s, \alpha\}$:

$$\begin{aligned} \boldsymbol{\alpha}[\tau] &= \arg\min_{\alpha[\tau]} \mathcal{A}(\alpha[\tau] \mid s[\tau]) \\ &\Rightarrow \delta_{\boldsymbol{\alpha}[\tau]} \mathcal{A}(\boldsymbol{\alpha}[\tau] \mid s[\tau]) = 0 \\ &\Rightarrow \dot{\boldsymbol{\alpha}} = (Q_{\alpha\alpha} - \Gamma_{\alpha\alpha})\nabla_\alpha F(s, \boldsymbol{\alpha}) \end{aligned} \tag{1.3}$$

This means the most likely path conforms to a variational principle of least action, where variational free energy is an upper bound on surprisal:

$$\begin{aligned} F(\pi) &\triangleq \underbrace{E_q[\Im(\eta, s, \alpha)]}_{\text{Energy}} - \underbrace{H[q_\mu(\eta)]}_{\text{Entropy}} \\ &= \underbrace{\Im(s,\alpha)}_{\text{Surprisal}} + \underbrace{D[q_\mu(\eta) \,\|\, p(\eta \mid s, \alpha)]}_{\text{Divergence}} \\ &= \underbrace{E_q[\Im(s, \alpha \mid \eta)]}_{\text{Inaccuracy}} + \underbrace{D[q_\mu(\eta) \,\|\, p(\eta)]}_{\text{Complexity}} \geq \Im(s,\alpha) \end{aligned} \tag{1.4}$$

This functional can be expressed in several forms; namely, an energy minus the entropy of the variational density, which is equivalent to the surprise associated with systemic states (i.e., *surprisal*) plus the KL divergence between the variational and posterior density (i.e., *divergence*). In turn, this can be decomposed into the negative log likelihood of systemic states (i.e., *inaccuracy*) and the KL divergence between posterior and prior densities (i.e., *complexity*).

Proof: the most likely trajectory – that minimises action – obtains when the random fluctuations about the flow take their most likely value of zero. By (1.2) the flow of the most likely autonomous states $\boldsymbol{\alpha} = \{\mathbf{a}, \boldsymbol{\mu}\}$ can be expressed as a gradient flow on surprisal or, by definition, variational free energy:

$$\begin{aligned} \boldsymbol{\alpha}[\tau] &= \arg\min_{\alpha[\tau]} \mathcal{A}(\alpha[\tau] \mid s[\tau]) \Rightarrow \\ \dot{\boldsymbol{\alpha}} &= (Q_{\alpha\alpha} - \Gamma_{\alpha\alpha}) \nabla_{\alpha} \Im(s, \boldsymbol{\alpha}) \\ &= (Q_{\alpha\alpha} - \Gamma_{\alpha\alpha}) \nabla_{\alpha} F(s, \boldsymbol{\alpha}) \end{aligned} \tag{1.5}$$

Where, for the most likely internal state, $\boldsymbol{\mu} \in \boldsymbol{\alpha}$:

$$F(s, \boldsymbol{\alpha}) = \Im(s, \boldsymbol{\alpha}) + \underbrace{D[q_{\boldsymbol{\mu}}(\eta) \,||\, p(\eta \mid s, \boldsymbol{\alpha})]}_{Divergence} = \Im(s, \boldsymbol{\alpha}) \tag{1.6}$$

The equivalence between variational free energy and the surprisal of systemic states follows from the definition of the variational density that renders the divergence zero □

Given this stipulative formulation of gradient flows under a Markov blanket, one can now use the path integral formalism to characterise the most likely path of autonomous states from any initial state.

Corollary (path integral formulation). *Under some simplifying assumptions, the action of autonomous paths from any initial systemic state is upper bounded by expected free energy*:

$$\mathcal{A}(\alpha[\tau]\,|\,\pi_0) \leq G(\alpha[\tau]) \tag{1.7}$$

Expected free energy is defined as:

$$\begin{aligned} G(\alpha[\tau]) &\triangleq \underbrace{E_q[\Im(\eta, s, \alpha_\tau)]}_{\textit{Energy}} - \underbrace{H[q_\tau(\eta)]}_{\textit{Entropy}} \\ &= \underbrace{E_q[\Im(s, \alpha_\tau)]}_{\textit{Expected surprisal}} + \underbrace{D[q_\tau(\eta\,|\,s) \,\|\, p(\eta\,|\,s, \alpha_\tau)]]}_{\textit{Expected divergence}} - \underbrace{D[q_\tau(\eta\,|\,s) \,\|\, q_\tau(\eta)]}_{\textit{Information gain}} \\ &= \underbrace{E_q[\Im(s, \alpha_\tau\,|\,\eta)]}_{\textit{Ambiguity}} + \underbrace{D[q_\tau(\eta) \,\|\, p(\eta)]}_{\textit{Risk}} \\ &\geq \mathcal{A}(\alpha[\tau]\,|\,\pi_0) \end{aligned} \tag{1.8}$$

The expectation in (1.8) is under the predictive density over hidden and sensory states, conditioned upon the initial systemic state and subsequent trajectory of autonomous states:

$$q_\tau(s, \eta) \triangleq p(s, \eta, \tau\,|\,\alpha[\tau], \pi_0) \tag{1.9}$$

The expected free energy in (1.8) has been formulated to emphasise the formal correspondence with variational free energy in (1.4): where the complexity and accuracy terms become *risk* (i.e., expected complexity) and *ambiguity* (i.e., expected inaccuracy).

In summary, variational free energy is an upper bound on the surprisal of systemic states– and expected free energy is an upper bound on the action of autonomous states. On a conceptual note, the role of nonequilibrium steady-state takes on a different aspect, depending upon whether the variational dynamics above are thought of in terms of gradient flows (i.e., the variational free energy lemma) or as picking out the most likely paths (i.e., the path integral corollary).

From the point of view of a statistician, the gradient flow formulation regards the probability density at nonequilibrium steady-state as a generative model; in other words, a probabilistic specification of the sensory impressions of external states hidden behind the Markov blanket.

It is this dynamics that licenses an interpretation of self-organisation in terms of statistical (i.e., approximate Bayesian) inference.

The picture changes when we consider the path integral formulation. Here, we are picking out trajectories of autonomous states (i.e., active and internal states) that are most likely under the generative model. On this view, the generative model can be regarded as some prior beliefs about the sensory states (and their external causes) that will be encountered in the future. In other words, the generative model prescribes the attracting set that the system will autonomously work towards – by apparently selecting the paths of activity that lead to these attracting states. This enactive perspective makes it look as if the generative model is no longer simply an explanation for sensory samples but a specification of the states a system aspires to.

Acknowledgements:

We thank Paul Badcock, Shaun Gallagher, Casper Hesp, Dan Hutto, Safae Essafi, Michael Kirchhoff, Sander Van de Cruys, Alan Jürgens, Thomas Parr, Ian Robertson, Ryan Smith, Anna Strasser, Auguste Nahas, Erik Rietveld, Jonathan St-Onge, Simon Tremblay, Jared Vasil, and Eric White, Julian Xue, and all those present at the Naturally Evolving Minds conference at University of Wollongong (20-23 February 2018) for helpful discussions and comments. We also sincerely thank the editor, Barbara Finley, and the anonymous reviewers who provided us with valuable feedback.

Funding

This research was produced thanks in part to funding from the Canada First Research Excellence Fund, awarded to McGill University for the Healthy Brains for Healthy Lives

initiative (S. P. L. Veissière and M. J. D. Ramstead), as well as a Postgraduate Research Scholarship in the Philosophy of Biomedicine (A. Constant – Ref: SC2730) as part of the ARC Australian Laureate Fellowship project *A Philosophy of Medicine for the 21st Century* (Ref: FL170100160) (A. Constant), a Joseph-Armand Bombardier Canada Doctoral Scholarship and a Michael Smith Foreign Study Supplements award from the Social Sciences and Humanities Research Council of Canada (M. J. D. Ramstead), and by a Wellcome Principal Research Fellowship (K. J. Friston – Ref: 088130/Z/09/Z).

References

Adams, Rick A., Quentin J. M. Huys, and Jonathan P. Roiser. 2016. "Computational Psychiatry: Towards a Mathematically Informed Understanding of Mental Illness." *Journal of Neurology, Neurosurgery, and Psychiatry* 87 (1): 53–63.

Andrews, Paul W., Steven W. Gangestad, and Dan Matthews. 2002. "Adaptationism--How to Carry out an Exaptationist Program." *The Behavioral and Brain Sciences* 25 (4): 489–504; discussion 504–53.

Apperly, Ian A., and Stephen A. Butterfill. 2009. "Do Humans Have Two Systems to Track Beliefs and Belief-like States?" *Psychological Review* 116 (4): 953–70.

Asch, Solomon E. 1956. "Studies of Independence and Conformity: I. A Minority of One against a Unanimous Majority." Psychological Monographs: General and Applied 70 (9): 1.

Astuti, Rita, and Maurice Bloch. 2015. "The Causal Cognition of Wrong Doing: Incest, Intentionality, and Morality." *Frontiers in Psychology* 6 (February): 136.

Badcock, Paul Benjamin. 2012. "Evolutionary Systems Theory: A Unifying Meta-Theory of Psychological Science." *Review of General Psychology: Journal of Division 1, of the American Psychological Association* 16 (1): 10–23.

Badcock, Paul Benjamin, Christopher G. Davey, Sarah Whittle, Nicholas B. Allen, and Karl J. Friston. 2017. "The Depressed Brain: An Evolutionary Systems Theory." *Trends in Cognitive Sciences* 21 (3): 182–94.

Badcock, P. B., Friston, K. J., & Ramstead, M. J. (2019). The hierarchically mechanistic mind: A free-energy formulation of the human psyche. Physics of life Reviews.

Baldwin, Mark W. 1992. "Relational Schemas and the Processing of Social Information." *Psychological Bulletin* 112 (3): 461.

Bengio, Yoshua. 2014. "Evolving Culture Versus Local Minima." In *Growing Adaptive Machines*, 109–38. Studies in Computational Intelligence. Springer, Berlin, Heidelberg.

Bertolotti, Tommaso, and Lorenzo Magnani. 2017. "Theoretical Considerations on Cognitive Niche Construction." *Synthese* 194 (12): 4757–79.

Berwick, R. C., and N. Chomsky. 2013. "Poverty of the Stimulus Stands: Why Recent Challenges Fail." *Rich Languages from*.

Bijleveld, Erik, Daan Scheepers, and Naomi Ellemers. 2012. "The Cortisol Response to Anticipated Intergroup Interactions Predicts Self-Reported Prejudice." *PloS One* 7 (3): e33681.

Bloom, Paul. 2005. *Descartes' Baby: How the Science of Child Development Explains What Makes Us Human*. Random House.

———. 2017. *Against Empathy: The Case for Rational Compassion*. Edited by Random House.

Bourdieu, Pierre. 1977. *Equisse D'une Théorie de La Pratique*. Cambridge University Press.

———. 1984. *Distinction: A Social Critique of the Judgement of Taste*. Harvard University Press.

Bolis, Dimitris, Joshua Balsters, Nicole Wenderoth, Cristina Becchio, and Leonhard Schilbach. 2017. "Beyond Autism: Introducing the Dialectical Misattunement

Hypothesis and a Bayesian Account of Intersubjectivity." Psychopathology 50 (6). https://doi.org/10.1159/000484353.

Bolis, Dimitris, and Leonhard Schilbach. 2017. "Observing and Participating in Social Interactions: Action Perception and Action Control across the Autistic Spectrum." Developmental Cognitive Neuroscience, January. https://doi.org/10.1016/j.dcn.2017.01.009.

Boyd, Robert, and Peter J. Richerson. 2005. *The Origin and Evolution of Cultures*. Oxford University Press.

Boyer, Pascal. 2018. *Minds Make Societies: How Cognition Explains the World Humans Create*. Yale University Press.

Brown, Donald E. 2004. "Human Universals, Human Nature & Human Culture." Daedalus 133 (4): 47–54. https://doi.org/10.1162/0011526042365645.

Bruineberg, Jelle, and Erik Rietveld. 2014. "Self-Organization, Free Energy Minimization, and Optimal Grip on a Field of Affordances." *Frontiers in Human Neuroscience* 8 (August): 599.

Bruineberg, Jelle, Erik Rietveld, Thomas Parr, Leendert van Maanen, and Karl J. Friston. 2018. "Free-Energy Minimization in Joint Agent-Environment Systems: A Niche Construction Perspective." *Journal of Theoretical Biology* 455 (October): 161–78.

Burkart, Judith M., Sarah B. Hrdy, and Carel P. Van Schaik. 2009. "Cooperative Breeding and Human Cognitive Evolution." *Evolutionary Anthropology: Issues, News, and Reviews* 18 (5): 175–86.

Carruthers, Peter, and Peter K. Smith. 1996. *Theories of Theories of Mind*. Cambridge University Press.

Chemero, Anthony. 2009. "Radical Embodied Cognition." Cambridge, MA: MIT Press.

Cheng, Joey T., Jessica L. Tracy, Tom Foulsham, Alan Kingstone, and Joseph Henrich. 2013.

"Two Ways to the Top: Evidence That Dominance and Prestige Are Distinct yet Viable Avenues to Social Rank and Influence." *Journal of Personality and Social Psychology* 104 (1): 103–25.

Chomsky, Noam. 1996. *Studies on Semantics in Generative Grammar*. Walter de Gruyter.

Christensen, W., & Michael, J. (2016). From two systems to a multi-systems architecture for mindreading. *New Ideas in Psychology*, 40, 48-64.

Chudek, Maciej, Rita McNamara, Susan Burch, Paul Bloom, and Joseph Henrich. 2013. "Developmental and Cross-Cultural Evidence for Intuitive Dualism." *Psychological Science* 20.

Cialdini, Robert B., and Noah J. Goldstein. 2004. "Social Influence: Compliance and Conformity." *Annual Review of Psychology* 55: 591–621.

Clark, Andy. 2006. "Language, Embodiment, and the Cognitive Niche." *Trends in Cognitive Sciences* 10 (8): 370–74.

———. 2008. *Supersizing the Mind: Embodiment, Action, and Cognitive Extension*. OUP USA.

———. 2013. "Whatever next? Predictive Brains, Situated Agents, and the Future of Cognitive Science." *The Behavioral and Brain Sciences* 36 (03): 181–204.

Clark, Andy, and David Chalmers. 1998. "The Extended Mind." *Analysis* 58 (1): 7–19.

Clark, Kenneth B. 1988. *Prejudice and Your Child*. Wesleyan University Press.

Clark, Kenneth B., and Mamie K. Clark. 1939. "The Development of Consciousness of Self and the Emergence of Racial Identification in Negro Preschool Children." *The Journal of Social Psychology* 10 (4): 591–99.

Coopersmith, Jennifer. 2017. The Lazy Universe: An Introduction to the Principle of Least Action. Oxford University Press.

Constant, Axel., Maxwell J. D. Ramstead, Samuel P. L. Veissière, and Karl J. Friston. 2019.

"Regimes of Expectations: An Active Inference Model of Social Conformity and Decision Making." Frontiers in Psychology. https://doi.org/10.3389/fpsyg.2019.00679.

Constant, Axel, J. Bervoets, K. Hens, and S. Van de Cruys. 2018a. "Precise Worlds for Certain Minds: An Ecological Perspective on the Relational Self in Autism." *Topoi. An International Review of Philosophy*, 1–13.

Constant, Axel, Maxwell J. D. Ramstead, Samuel P. L. Veissière, John O. Campbell, and K. Friston. 2018a. "A Variational Approach to Niche Construction." *Journal of the Royal Society, Interface / the Royal Society*.

Csibra, Gergely, and György Gergely. 2009. "Natural Pedagogy." *Trends in Cognitive Sciences* 13 (4): 148–53.

Cullen, Maell, Ben Davey, Karl J. Friston, and Rosalyn J. Moran. 2018. "Active Inference in OpenAI Gym: A Paradigm for Computational Investigations Into Psychiatric Illness." Biological Psychiatry. Cognitive Neuroscience and Neuroimaging 3 (9): 809–18. https://doi.org/10.1016/j.bpsc.2018.06.010.

———. 2011. "Natural Pedagogy as Evolutionary Adaptation." *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 366 (1567): 1149–57.

De Castro, E. V. 2009. Métaphysiques Cannibales: Lignes D'anthropologie Post-Structurale. Paris, France: Presses universitaires de France.

Dehaene, Stanislas, and Laurent Cohen. 2007. "Cultural Recycling of Cortical Maps." *Neuron* 56 (2): 384–98.

Dijk, Ludger van, and Erik Rietveld. 2016. "Foregrounding Sociomaterial Practice in Our Understanding of Affordances: The Skilled Intentionality Framework." *Frontiers in Psychology* 7: 1969.

Dunbar, R. I. M. 2003. "The Social Brain: Mind, Language, and Society in Evolutionary Perspective." *Annual Review of Anthropology* 32 (1): 163–81.

———. 2004. “Gossip in Evolutionary Perspective.” *Review of General Psychology: Journal of Division 1, of the American Psychological Association* 8 (2): 100–110.

Duranti, Alessandro. 2015. *The Anthropology of Intentions*. Cambridge University Press.

Durkheim, Emile. 1985/2014. *The Rules of Sociological Method: And Selected Texts on Sociology and Its Method*. Simon and Schuster.

Dweck, C. S. 2013. “Self-Theories: Their Role in Motivation, Personality, and Development.” https://content.taylorfrancis.com/books/download?dac=C2009-0-07336-6&isbn=9781317710332&format=googlePreviewPdf.

Einarsson, Anna, and Tom Ziemke. 2017. “Exploring the Multi-Layered Affordances of Composing and Performing Interactive Music with Responsive Technologies.” *Frontiers in Psychology* 8 (September): 1701.

Fabry, Regina E. 2017. “Betwixt and between: The Enculturated Predictive Processing Approach to Cognition.” *Synthese*, 1–36.

Feinman, Saul. 1982. “Social Referencing in Infancy.” Merrill-Palmer Quarterly, 445–70. https://www.jstor.org/stable/pdf/23086154.pdf

Feldman, Harriet, and Karl J. Friston. 2010. “Attention, Uncertainty, and Free-Energy.” *Frontiers in Human Neuroscience* 4 (December): 215.

Feldman, Jacob. 2013. “Tuning Your Priors to the World.” *Topics in Cognitive Science* 5 (1): 13–34.

Friston, Karl J. 2005. “A Theory of Cortical Responses.” *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 360 (1456): 815–36.

———. 2010. “The Free-Energy Principle: A Unified Brain Theory?” *Nature Reviews. Neuroscience* 11 (2): 127–38.

———. 2011. “Embodied Inference: Or I Think Therefore I Am, If I Am What I Think.” *The Implications of Embodiment (Cognition and Communication)*, 89–125.

———. 2013. “Life as We Know It.” *Journal of the Royal Society, Interface / the Royal Society* 10 (86): 20130475.

Friston, Karl J., Philipp Schwartenbeck, Thomas FitzGerald, Michael Moutoussis, Timothy Behrens, and Raymond J. Dolan. 2014. “The Anatomy of Choice: Dopamine and Decision-Making.” Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences 369 (1655). https://doi.org/10.1098/rstb.2013.0481.

Friston, Karl J., Thomas FitzGerald, Francesco Rigoli, Philipp Schwartenbeck, John O Doherty, and Giovanni Pezzulo. 2016. “Active Inference and Learning.” *Neuroscience and Biobehavioral Reviews* 68 (September): 862–79.

Friston, Karl J., Thomas FitzGerald, Francesco Rigoli, Philipp Schwartenbeck, and Giovanni Pezzulo. 2017. “Active Inference: A Process Theory.” *Neural Computation* 29 (1): 1–49.

Friston, Karl J., James Kilner, and Lee Harrison. 2006. “A Free Energy Principle for the Brain.” *Journal of Physiology, Paris* 100 (1-3): 70–87.

Friston, Karl J., Marco Lin, Christopher D. Frith, Giovanni Pezzulo, J. Allan Hobson, and Sasha Ondobaka. 2017. “Active Inference, Curiosity and Insight.” *Neural Computation* 29 (10): 2633–83.

Friston, Karl J., and Klaas E. Stephan. 2007. “Free-Energy and the Brain.” *Synthese* 159 (3): 417–58.

Friston, Karl J., Klaas Enno Stephan, Read Montague, and Raymond J. Dolan. 2014. “Computational Psychiatry: The Brain as a Phantastic Organ.” *The Lancet Psychiatry* 1 (2): 148–58.

Friston, Karl, Francesco Rigoli, Dimitri Ognibene, Christoph Mathys, Thomas Fitzgerald, and Giovanni Pezzulo. 2015. “Active Inference and Epistemic Value.” *Cognitive Neuroscience* 6 (4): 187–214.

Friston, Karl, Christopher Thornton, and Andy Clark. 2012. “Free-Energy Minimization and

the Dark-Room Problem." *Frontiers in Psychology* 3. https://doi.org/10.3389/fpsyg.2012.00130.

Fulda, Fermín C. 2017. "Natural Agency: The Case of Bacterial Cognition." *Journal of the American Philosophical Association* 3 (1): 69–90.

Gallagher, Shaun. 2017. *Enactivist Interventions: Rethinking the Mind*. Oxford University Press.

Gallagher, Shaun, and Micah Allen. 2016. "Active Inference, Enactivism and the Hermeneutics of Social Cognition." *Synthese*, 1–22.

Gallese, V., and A. Goldman. 1998. "Mirror Neurons and the Simulation Theory of Mind-Reading." *Trends in Cognitive Sciences* 2 (12): 493–501.

Gavrilets, Sergey, and Aaron Vose. 2006. "The Dynamics of Machiavellian Intelligence." *Proceedings of the National Academy of Sciences of the United States of America* 103 (45): 16823–28.

Geertz, Clifford. 1973. *The Interpretation of Cultures*. Vol. 5043. Basic books.

Gibson, James Jerome. 1979. *The Ecological Approach to Visual Perception*. Houghton Mifflin.

Goffman, Erving. 2009. *Relations in Public*. Transaction Publishers.

Gold, Joel, and Ian Gold. 2015. *Suspicious Minds : How Culture Shapes Madness*. New York: Free Press.

Goldman, Alvin. 2006. *Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading*. Oxford University Press, USA.

Goldstein, Julie, Jules Davidoff, and Debi Roberson. 2009. "Knowing Color Terms Enhances Recognition: Further Evidence from English and Himba." *Journal of Experimental Child Psychology* 102 (2): 219–38.

Gopnik, Alison, and Henry M. Wellman. 2012. "Reconstructing Constructivism: Causal

Models, Bayesian Learning Mechanisms, and the Theory Theory." *Psychological Bulletin* 138 (6): 1085–1108.

Hacking, Ian. 1998. *Mad Travelers: Reflections on the Reality of Transient Mental Illnesses*. University of Virginia Press.

Hamilton, Antonia F. de C. 2008. "Emulation and Mimicry for Social Interaction: A Theoretical Approach to Imitation in Autism." Quarterly Journal of Experimental Psychology 61 (1): 101–15. https://doi.org/10.1080/17470210701508798.

Heine, Steven J., Timothy Takemoto, Sophia Moskalenko, Jannine Lasaleta, and Joseph Henrich. 2008. "Mirrors in the Head: Cultural Variation in Objective Self-Awareness." *Personality & Social Psychology Bulletin* 34 (7): 879–87.

Henrich, Joseph. 2015. *The Secret of Our Success: How Culture Is Driving Human Evolution, Domesticating Our Species, and Making Us Smarter*. Princeton, NJ: Princeton University Press.

Henrich, Joseph, and F. J. Gil-White. 2001. "The Evolution of Prestige: Freely Conferred Deference as a Mechanism for Enhancing the Benefits of Cultural Transmission." *Evolution and Human Behavior: Official Journal of the Human Behavior and Evolution Society* 22 (3): 165–96.

Hewlett, Barry S. 1994. *Intimate Fathers : The Nature and Context of Aka Pygmy Paternal Infant Care*. Ann Arbor, Mich.: University of Michigan Press.

Hewlett, B. S. 2017. *Hunter-Gatherer Childhoods: Evolutionary, Developmental, and Cultural Perspectives*. Routledge.

Heyes, Cecilia. 2018. *Cognitive Gadgets: The Cultural Evolution of Thinking*. Harvard University Press.

Heyes, Cecilia M., and Chris D. Frith. 2014. "The Cultural Evolution of Mind Reading." *Science* 344 (6190): 1243091.

Hohwy, Jakob. 2013. *The Predictive Mind*. Oxford: Oxford University Press.

Howes, David. 2011. "Reply to Tim Ingold." *Social Anthropology* 19 (3): 318–22.

Hrdy, Sarah Blaffer. 2011. *Mothers and Others*. Harvard University Press.

Huneman, Philippe, and Edouard Machery. 2015. "Evolutionary Psychology: Issues, Results, Debates." In *Handbook of Evolutionary Thinking in the Sciences*, edited by Thomas Heams, Philippe Huneman, Guillaume Lecointre, and Marc Silberstein, 647–57. Dordrecht: Springer Netherlands.

Hutto, Daniel D. 2012a. *Folk Psychological Narratives: The Sociocultural Basis of Understanding Reasons*. MIT Press.

Hutto, Daniel D., Michael D. Kirchhoff, and Erik Myin. 2014. "Extensive Enactivism: Why Keep It All In?" *Frontiers in Human Neuroscience* 8: 706.

Hutto, Daniel, and Erik Myin. 2013. "Radical Enactivism: Basic Minds without Content." Cambridge, MA: MIT Press.

———. 2017. *Evolving Enactivism: Basic Minds Meet Content*. MIT Press.

Hutto, Daniel, and Glenda Satne. 2015. "The Natural Origins of Content." *Philosophia* 43 (3): 521–36.

Huys, Quentin J. M., Tiago V. Maia, and Michael J. Frank. 2016. "Computational Psychiatry as a Bridge from Neuroscience to Clinical Applications." *Nature Neuroscience* 19 (3): 404–13.

Ignatow, G. 2009. "Why the Sociology of Morality Needs Bourdieu's Habitus." *Sociological Inquiry*. http://onlinelibrary.wiley.com/doi/10.1111/j.1475-682X.2008.00273.x/full.

Ingold, Tim. 2001. "From the Transmission of Representations to the Education of Attention." *The Debated Mind: Evolutionary Psychology versus Ethnography*, 113–53.

———. 2016. *Lines: A Brief History*. Routledge.

Jack, Anthony I. 2014. "A Scientific Case for Conceptual Dualism: The Problem of

Consciousness and the Opposing Domains Hypothesis." Oxford Studies in Experimental Philosophy 1: 1–32.

Johnson, Susan C., Carol S. Dweck, and Frances S. Chen. 2007. "Evidence for Infants' Internal Working Models of Attachment." *Psychological Science* 18 (6): 501–2.

Joffily, Mateus, and Giorgio Coricelli. 2013. "Emotional Valence and the Free-Energy Principle." PLoS Computational Biology 9 (6): e1003094. https://doi.org/10.1371/journal.pcbi.1003094.

Kahneman, Daniel. 2011. *Thinking, Fast and Slow*. Macmillan.

Kaplan, Raphael, and Karl J. Friston. 2018. "Planning and Navigation as Active Inference." *Biological Cybernetics*, March. https://doi.org/10.1007/s00422-018-0753-2.

Kaufmann, Laurence, and Fabrice Clément. 2014. "Wired for Society: Cognizing Pathways to Society and Culture." *Topoi. An International Review of Philosophy* 33 (2): 459–75.

Keane, Webb. 2015. "'Varieties of Ethical Stance.'" In *Four Lectures on Ethics: Anthropological Perspectives*, edited by Lambek, Michael, Veena Das, Didier Fassin, and Webb Keane. Chicago: Hau.

Kelly, Daniel, Luc Faucher, and Edouard Machery. 2010. "Getting Rid of Racism: Assessing Three Proposals in Light of Psychological Evidence: Getting Rid of Racism." *Journal of Social Philosophy* 41 (3): 293–322.

Kiebel, S. J., & Friston, K. J. (2011). Free energy and dendritic self-organization. Frontiers in systems neuroscience, 5, 80.

Kiebel, Stefan J., Jean Daunizeau, and Karl J. Friston. 2008. "A Hierarchy of Time-Scales and the Brain." *PLoS Computational Biology* 4 (11): e1000209.

Kinzler, Katherine D., Emmanuel Dupoux, and Elizabeth S. Spelke. 2007. "The Native Language of Social Cognition." *Proceedings of the National Academy of Sciences of the United States of America* 104 (30): 12577–80.

Kinzler, Katherine D., and Elizabeth S. Spelke. 2011. "Do Infants Show Social Preferences for People Differing in Race?" *Cognition* 119 (1): 1–9.

Kirmayer, Laurence J. 1989. "Cultural Variations in the Response to Psychiatric Disorders and Emotional Distress." *Social Science & Medicine* 29 (3): 327–39.

Kirmayer, Laurence J., and Ian Gold. 2011. "Re-Socializing Psychiatry: Critical Neuroscience and the Limits of Reductionism." In *Critical Neuroscience*, edited by Suparna Choudhury and Jan Slaby, 185:305–30. Oxford, UK: Wiley-Blackwell.

Kirmayer, Laurence J., Ana Gomez-Carrillo, and Samuel P. L. Veissière. 2017. "Culture and Depression in Global Mental Health: An Ecosocial Approach to the Phenomenology of Psychiatric Disorders." *Social Science & Medicine* 183 (June): 163–68.

Kirmayer, Laurence J., Robert Lemelson, and Constance A. Cummings. 2015. *Re-Visioning Psychiatry: Cultural Phenomenology, Critical Neuroscience, and Global Mental Health*. Cambridge University Press.

Kirmayer, Laurence J. 2015. "Re-Visioning Psychiatry: Toward an Ecology of Mind in Health and Illness." In Cultural Phenomenology, Critical Neuroscience and Global Mental Health, edited by Laurence J. Kirmayer, R. Robert Lemelson, and Constance A. Cummings, 622–60. Cambridge: Cambridge University Press.

Kirmayer, Laurence J., and Maxwell J. D. Ramstead. 2017. "Embodiment and Enactment in Cultural Psychiatry." In *Embodiment, Enaction, and Culture: Investigating the Constitution of the Shared World*. MIT Press.

Kirmayer, Laurence J., and A. Young. 1998. "Culture and Somatization: Clinical, Epidemiological, and Ethnographic Perspectives." *Psychosomatic Medicine* 60 (4): 420–30.

Kiverstein, Julian, Mark Miller, and Erik Rietveld. 2017. "The Feeling of Grip: Novelty, Error Dynamics, and the Predictive Brain." *Synthese*, October. https://doi.org/10.1007/s11229-017-1583-9.

Kurzban, R., and S. Neuberg. 2005. “Managing Ingroup and Outgroup Relationships.” In *The Handbook of Evolutionary Psychology*, edited by D. M. Buss, 653–75. Hoboken, NJ, US: John Wiley & Sons Inc.

Lakoff, George, and Mark Johnson. 1980. “The Metaphorical Structure of the Human Conceptual System.” *Cognitive Science* 4 (2): 195–208.

Laland, Kevin N. 2018. Darwin’s Unfinished Symphony: How Culture Made the Human Mind. Princeton University Press.

Laland, Kevin N., Tobias Uller, Marcus W. Feldman, Kim Sterelny, Gerd B. Müller, Armin Moczek, Eva Jablonka, and John Odling-Smee. 2015. “The Extended Evolutionary Synthesis: Its Structure, Assumptions and Predictions.” *Proceedings. Biological Sciences / The Royal Society* 282 (1813): 20151019.

Lebois, Lauren A. M., Christine D. Wilson-Mendenhall, W. Kyle Simmons, Lisa Feldman Barrett, and Lawrence W. Barsalou. 2018. “Learning Situated Emotions.” *Neuropsychologia*, January.

Lévy, Robert. 1984. “Emotion, Knowing and Culture.” In *Culture Theory: Essays on Mind, Self, and Emotion*, edited by R. Shweder and R. LeVine, 214–37. Cambridge: Cambridge University Press.

Levy, Robert I. 1975. *Tahitians: Mind and Experience in the Society Islands*. University of Chicago Press.

Luhrmann, Tanya. 2011. “Toward an Anthropological Theory of Mind.” *Suomen Antropologi: Journal of the Finnish Anthropological Society* 36 (4): 5–69.

Luo, Siyang, Bingfeng Li, Yina Ma, Wenxia Zhang, Yi Rao, and Shihui Han. 2015. “Oxytocin Receptor Gene and Racial Ingroup Bias in Empathy-Related Brain Activity.” *NeuroImage* 110 (April): 22–31.

Luo, Yuyan, and Renée Baillargeon. 2005. “Can a Self-Propelled Box Have a Goal?

Psychological Reasoning in 5-Month-Old Infants." *Psychological Science* 16 (8): 601–8.

Lühe, T. von der, V. Manera, I. Barisic, C. Becchio, K. Vogeley, and L. Schilbach. 2016. "Interpersonal Predictive Coding, Not Action Perception, Is Impaired in Autism." Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences 371 (1693). https://doi.org/10.1098/rstb.2015.0373.

Lutz, Antoine, Julie Brefczynski-Lewis, Tom Johnstone, and Richard J. Davidson. 2008. "Regulation of the Neural Circuitry of Emotion by Compassion Meditation: Effects of Meditative Expertise." *PloS One* 3 (3): e1897.

Machery, E. 2016. "De-Freuding Implicit Attitudes." *Implicit Bias and Philosophy*.

Machery, Edouard, and Luc Faucher. 2017. "Why Do We Think Racially? Culture, Evolution, and Cognition." In *Handbook of Categorization in Cognitive Science (Second Edition)*, edited by Henri Cohen and Claire Lefebvre, 1135–75. San Diego: Elsevier.

Madoka, Mayuzumi. 2003. "Haiku." In *Far Beyond the Field: Haiku by Japanese Women*, edited by Makoto Ueda. Columbia University Press.

Mahajan, Neha, and Amanda Woodward. 2009. "Seven-Month-Old Infants Selectively Reproduce the Goals of Animate But Not Inanimate Agents." *Infancy: The Official Journal of the International Society on Infant Studies* 14 (6): 667–79.

Malafouris, Lambros. 2015. "Metaplasticity and the Primacy of Material Engagement." *Time and Mind* 8 (4): 351–71.

Mameli, Matteo. 2001. "Mindreading, Mindshaping, and Evolution." *Biology and Philosophy* 16 (5): 595–626.

Mauss, Marcel. 1973. "Techniques of the Body." *Economy and Society* 2 (1): 70–88.

McCauley, Robert N., and Joseph Henrich. 2006. "Susceptibility to the Müller-Lyer Illusion, Theory-Neutral Observation, and the Diachronic Penetrability of the Visual Input System." *Philosophical Psychology* 19 (1): 79–101.

McGeer, Victoria. 2007. “The Regulative Dimension of Folk Psychology.” In Folk Psychology Re-Assessed, edited by Daniel D. Hutto and Matthew Ratcliffe, 137–56. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-1-4020-5558-4_8.

Menary, Richard. 2010. “The Extended Mind and Cognitive Integration.” In *The Extended Mind*, edited by Richard Menary. MIT Press.

Mercier, Hugo, and Dan Sperber. 2017a. *The Enigma of Reason*. Harvard University Press.

———. 2017b. *The Enigma of Reason*. Harvard University Press.

Michael, John, Wayne Christensen, and Søren Overgaard. 2014. “Mindreading as Social Expertise.” *Synthese* 191 (5): 817–40.

Milgram, S. 1963. “BEHAVIORAL STUDY OF OBEDIENCE.” *Journal of Abnormal Psychology* 67 (October): 371–78.

Miresco, Marc J., and Laurence J. Kirmayer. 2006. “The Persistence of Mind-Brain Dualism in Psychiatric Reasoning about Clinical Scenarios.” *The American Journal of Psychiatry* 163 (5): 913–18.

Mirza, M. Berk, Rick A. Adams, Christoph D. Mathys, and Karl J. Friston. 2016. “Scene Construction, Visual Foraging, and Active Inference.” Frontiers in Computational Neuroscience 10 (June): 56. https://doi.org/10.3389/fncom.2016.00056.

Montague, P. Read, Raymond J. Dolan, Karl J. Friston, and Peter Dayan. 2012. “Computational Psychiatry.” *Trends in Cognitive Sciences* 16 (1): 72–80.

Morgan, T. J. H., and K. N. Laland. 2012. “The Biological Bases of Conformity.” Frontiers in Neuroscience 6 (June): 87. https://doi.org/10.3389/fnins.2012.00087.

Navarrete, Carlos David, and Daniel M. T. Fessler. 2005. “Normative Bias and Adaptive Challenges: A Relational Approach to Coalitional Psychology and a Critique of Terror Management Theory.” *Evolutionary Psychology: An International Journal of Evolutionary Approaches to Psychology and Behavior* 3 (1): 147470490500300121.

Odling-Smee, John, Kevin N. Laland, and Marcus W. Feldman. 2003. *Niche Construction: The Neglected Process in Evolution*. Princeton University Press.

Olson, Kristina R., and Elizabeth S. Spelke. 2008. "Foundations of Cooperation in Young Children." *Cognition* 108 (1): 222–31.

Onishi, Kristine H., and Renée Baillargeon. 2005. "Do 15-Month-Old Infants Understand False Beliefs?" *Science* 308 (5719): 255–58.

Oudeyer, Pierre-Yves, and Frederic Kaplan. 2007. "What Is Intrinsic Motivation? A Typology of Computational Approaches." *Frontiers in Neurorobotics* 1 (November): 6.

Parr, Thomas, and Karl J. Friston. 2017a. "Uncertainty, Epistemics and Active Inference." *Journal of the Royal Society, Interface / the Royal Society* 14 (136). https://doi.org/10.1098/rsif.2017.0376.

———. 2017b. "Working Memory, Attention, and Salience in Active Inference." *Scientific Reports* 7 (1): 14678.

———. 2018. "Attention or Salience?" *Current Opinion in Psychology* 29 (October): 1–5.

Pauker, Kristin, Amanda Williams, and Jennifer R. Steele. 2016. "Children's Racial Categorization in Context." *Child Development Perspectives* 10 (1): 33–38.

Pezzulo, Giovanni, Emilio Cartoni, Francesco Rigoli, Léo Pio-Lopez, and Karl Friston. 2016. "Active Inference, Epistemic Value, and Vicarious Trial and Error." *Learning & Memory* 23 (7): 322–38.

Pezzulo, Giovanni, and Paul Cisek. 2016. "Navigating the Affordance Landscape: Feedback Control as a Process Model of Behavior and Cognition." *Trends in Cognitive Sciences* 20 (6): 414–24.

Phillips, M. L., A. W. Young, C. Senior, M. Brammer, C. Andrew, A. J. Calder, E. T. Bullmore, et al. 1997. "A Specific Neural Substrate for Perceiving Facial Expressions of Disgust." *Nature* 389 (6650): 495–98.

Pinker, S. 1999. “How the Mind Works.” *Annals of the New York Academy of Sciences* 882 (June): 119–27; discussion 128–34.

———. 2003. “Language as an Adaptation to the Cognitive Niche.” In *Language Evolution: States of the Art*, edited by S Kirby &, 16–37. New York: Oxford University Press.

Poerio, Giulia Lara, and Jonathan Smallwood. 2016. “Daydreaming to Navigate the Social World: What We Know, What We Don’t Know, and Why It Matters.” *Social and Personality Psychology Compass* 10 (11): 605–18.

Ramsey, William M. 2007. *Representation Reconsidered*. Cambridge University Press.

Ramstead, Maxwell J. D., Paul Benjamin Badcock, and Karl John Friston. 2017. “Answering Schrödinger’s Question: A Free-Energy Formulation.” *Physics of Life Reviews*, September. https://doi.org/10.1016/j.plrev.2017.09.001.

Ramstead, Maxwell J. D., Axel Constant, Paul B. Badcock, and K. Friston. 2018. “Variational Ecology and the Physics of Sentient Systems.” *Physics of Life Reviews*.

Ramstead, Maxwell J. D., Samuel P. L. Veissière, and Laurence J. Kirmayer. 2016. “Cultural Affordances: Scaffolding Local Worlds through Shared Intentionality and Regimes of Attention.” *Frontiers in Psychology* 7 (July): 1090.

Rietveld, Erik, and Anne Ardina Brouwers. 2017. “Optimal Grip on Affordances in Architectural Design Practices: An Ethnography.” *Phenomenology and the Cognitive Sciences* 16 (3): 545–64.

Rietveld, Erik, and Julian Kiverstein. 2014. “A Rich Landscape of Affordances.” *Ecological Psychology: A Publication of the International Society for Ecological Psychology* 26 (4): 325–52.

Robbins, Joel. 2008. “On Not Knowing Other Minds: Confession, Intention, and Linguistic Exchange in a Papua New Guinea Community.” *Anthropological Quarterly* 81 (2): 421–29.

Robbins, Joel, Julia Cassaniti, and T. M. Luhrmann. 2011. “The Constitution of Mind: What’s

in a Mind? Interiority and Boundedness." *Suomen Antropologi* 36 (4): 15–20.

Robbins, Joel, and Alan Rumsey. 2008. "Introduction: Cultural and Linguistic Anthropology and the Opacity of Other Minds." *Anthropological Quarterly* 81 (2): 407–20.

Roepstorff, Andreas, Jörg Niewöhner, and Stefan Beck. 2010. "Enculturing Brains through Patterned Practices." *Neural Networks: The Official Journal of the International Neural Network Society* 23 (8-9): 1051–59.

Rosaldo, Michelle Z. 1982. "The Things We Do with Words: Ilongot Speech Acts and Speech Act Theory in Philosophy." *Language In Society* 11 (2): 203–37.

Rozin, Paul, Jonathan Haidt, and Katrina Fincher. 2009. "Psychology. From Oral to Moral." *Science* 323 (5918): 1179–80.

Rumsey, Alan. 2013. "Intersubjectivity, Deception and the 'opacity of Other Minds': Perspectives from Highland New Guinea and beyond." *Language & Communication* 33 (3): 326–43.

Seligman, R., Choudhury, S., & Kirmayer, L. J. 2015. "Locating Culture in the Brain and in the World: From Social Categories to the Ecology of Mind." In Handbook of Cultural Neuroscience, edited by J. Chiao et al., 3–20. Oxford University Press.

Schilbach Leonhard. 2016. "Towards a Second-Person Neuropsychiatry." Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences 371 (1686): 20150081. https://doi.org/10.1098/rstb.2015.0081.

Schmidhuber, Jürgen. 2006. "Developmental Robotics, Optimal Artificial Curiosity, Creativity, Music, and the Fine Arts." *Connection Science* 18 (2): 173–87.

Schwartenbeck, P., & Friston, K. (2016). Computational phenotyping in psychiatry: a worked example. eneuro, 3(4).

Seth, Anil K., and Karl J. Friston. 2016. "Active Interoceptive Inference and the Emotional Brain." *Philosophical Transactions of the Royal Society of London. Series B, Biological*

Sciences 371 (1708).

Shapiro, Lawrence. 2010. *Embodied Cognition*. Routledge.

Shelley-Tremblay, J. F., and L. A. Rosén. 1996. "Attention Deficit Hyperactivity Disorder: An Evolutionary Perspective." *The Journal of Genetic Psychology* 157 (4): 443–53.

Spelke, Elizabeth S., and Katherine D. Kinzler. 2007. "Core Knowledge." *Developmental Science* 10 (1): 89–96.

Sperber, Dan. 1996. *Explaining Culture: A Naturalistic Approach*. Wiley.

———. 1997. "Intuitive and Reflective Beliefs." *Mind & Language* 12 (1): 67–83.

Stasch, Rupert. 2009. *Society of Others Kinship and Mourning in a West Papuan Place*. Berkeley: University of California Press.

Stephan, Klaas Enno, Lars Kasper, Lee M. Harrison, Jean Daunizeau, Hanneke E. M. den Ouden, Michael Breakspear, and Karl J. Friston. 2008. "Nonlinear Dynamic Causal Models for fMRI." *NeuroImage* 42 (2): 649–62.

Sterelny, Kim. 2012. *The Evolved Apprentice*. MIT Press.

Stotz, Karola. 2017. "Why Developmental Niche Construction Is Not Selective Niche Construction: And Why It Matters." *Interface Focus* 7 (5): 20160157.

Stotz, Karola, and Paul Edmund Griffiths. 2017. "A Developmental Systems Account of Human Nature." In *Why We Disagree About Human Nature*, edited by Tim Lewens and Elizabeth Hannon. Oxford & New York: Oxford University Press.

Stout, Dietrich, and Thierry Chaminade. 2007. "The Evolutionary Neuroscience of Tool Making." *Neuropsychologia* 45 (5): 1091–1100.

Stout, Dietrich, Nicholas Toth, Kathy Schick, and Thierry Chaminade. 2008. "Neural Correlates of Early Stone Age Toolmaking: Technology, Language and Cognition in Human Evolution." *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 363 (1499): 1939–49.

Sutton, John. 2010. “Exograms and Interdisciplinarity: History, the Extended Mind, and the Civilizing Process.” In *The Extended Mind*, edited by Richard Menary, 189–225. Cambridge, Mass.: MIT Press.

Swami, Viren, David A. Frederick, Toivo Aavik, Lidia Alcalay, Jüri Allik, Donna Anderson, Sonny Andrianto, et al. 2010. “The Attractive Female Body Weight and Female Body Dissatisfaction in 26 Countries across 10 World Regions: Results of the International Body Project I.” *Personality & Social Psychology Bulletin* 36 (3): 309–25.

Swanson, James, Robert Moyzis, John Fossella, Jin Fan, and Michael I. Posner. 2002. “Adaptationism and Molecular Biology: An Example Based on ADHD.” *The Behavioral and Brain Sciences* 25 (4): 530–31.

Taylor, Charles. 2016. *The Language Animal*. Harvard University Press.

Timmermans, Bert, Leonhard Schilbach, Antoine Pasquali, and Axel Cleeremans. 2012. “Higher Order Thoughts in Action: Consciousness as an Unconscious Re-Description Process.” Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences 367 (1594): 1412–23. https://doi.org/10.1098/rstb.2011.0421.

Tomasello, Michael. 2009. *Why We Cooperate*. MIT Press.

———. 2014. *A Natural History of Human Thinking*. Cambridge, MA: Harvard University Press.

Tomasello, Michael, Malinda Carpenter, Josep Call, Tanya Behne, and Henrike Moll. 2005. “Understanding and Sharing Intentions: The Origins of Cultural Cognition.” *The Behavioral and Brain Sciences* 28 (5): 675–91; discussion 691–735.

Tovo-Rodrigues, Luciana, Luis A. Rohde, Ana M. B. Menezes, Guilherme V. Polanczyk, Christian Kieling, Julia P. Genro, Luciana Anselmi, and Mara H. Hutz. 2013. “DRD4 Rare Variants in Attention-Deficit/Hyperactivity Disorder (ADHD): Further Evidence from a Birth Cohort Study.” *PloS One* 8 (12): e85164.

Trivers, R. 2000. “The Elements of a Scientific Theory of Self-Deception.” *Annals of the New York Academy of Sciences* 907 (April): 114–31.

Tschacher, Wolfgang, and Hermann Haken. 2007. “Intentionality in Non-Equilibrium Systems? The Functional Aspects of Self-Organized Pattern Formation.” *New Ideas in Psychology* 25 (1): 1–15.

Tybur, Joshua M., Debra Lieberman, Robert Kurzban, and Peter DeScioli. 2013. “Disgust: Evolved Function and Structure.” *Psychological Review* 120 (1): 65–84.

Veissière, Samuel P. L. 2016. “Varieties of Tulpa Experiences: The Hypnotic Nature of Human Sociality, Personhood, and Interphenomenality.” *Hypnosis and Meditation: Towards an Integrative Science of Conscious Planes*, 55–76.

———. 2017. “Cultural Markov Blankets? Mind the Other Minds Gap!: Comment on ‘Answering Schrödinger’s Question: A Free-Energy Formulation.’” *Physics of Life Reviews*, November.

Wang, Yong-Guang, Yi-Qiang Wang, Shu-Lin Chen, Chun-Yan Zhu, and Kai Wang. 2008. “Theory of Mind Disability in Major Depression with or without Psychotic Symptoms: A Componential View.” *Psychiatry Research* 161 (2): 153–61.

Whiten, Andrew, and David Erdal. 2012. “The Human Socio-Cognitive Niche and Its Evolutionary Origins.” *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 367 (1599): 2119–29.

Williams, Bernard. 2011. *Ethics and the Limits of Philosophy*. Taylor & Francis.

Wright, Len Tiu, Clive Nancarrow, and Pamela M. H. Kwok. 2001. “Food Taste Preferences and Cultural Influences on Consumption.” *British Food Journal* 103 (5): 348–57.

Zatzick, D. F., and J. E. Dimsdale. 1990. “Cultural Variations in Response to Painful Stimuli.” *Psychosomatic Medicine* 52 (5): 544–57.

Zawidzki, Tadeusz W. 2008. “The Function of Folk Psychology: Mind Reading or Mind

Shaping?" *Philosophical Explorations: An International Journal for the Philosophy of Mind and Action* 11 (3): 193–210.

Zawidzki, Tadeusz Wieslaw. 2013. *Mindshaping: A New Framework for Understanding Human Social Cognition*. MIT Press.