frontiers in Systems Neuroscience

HYPOTHESIS AND THEORY
published: 13 October 2021
doi: 10.3389/fnsys.2021.669810

Trust as Extended Control: Human-Machine Interactions as Active Inference

Felix Schoeller[1,2]*, *Mark Miller*[3,4], *Roy Salomon*[2] *and Karl J. Friston*[5]

[1] Massachusetts Institute of Technology, Cambridge, MA, United States, [2] Gonda Multidisciplinary Brain Research Center, Bar-Ilan University, Ramat Gan, Israel, [3] Center for Human Nature, Artificial Intelligence and Neuroscience, Hokkaido University, Sapporo, Japan, [4] Department of Informatics, University of Sussex, Brighton, United Kingdom, [5] Wellcome Trust Centre for Neuroimaging, University College London, London, United Kingdom

Edited by:
Mikhail Lebedev,
Duke University, United States

Reviewed by:
Victor de Lafuente,
National Autonomous University of Mexico, Mexico
Alessandro Umbrico,
Institute of Cognitive Sciences and Technologies, Italian National Research Council, Italy
Roel Stephan Pieters,
Tampere University of Technology, Finland
Andrea Orlandini,
National Research Council (CNR), Italy

***Correspondence:**
Felix Schoeller
felixsch@mit.edu

Received: *19 February 2021*
Accepted: *16 August 2021*
Published: *13 October 2021*

Citation:
Schoeller F, Miller M, Salomon R and Friston KJ (2021) Trust as Extended Control: Human-Machine Interactions as Active Inference. Front. Syst. Neurosci. 15:669810. doi: 10.3389/fnsys.2021.669810

In order to interact seamlessly with robots, users must infer the causes of a robot's behavior–and be confident about that inference (and its predictions). Hence, trust is a necessary condition for human-robot collaboration (HRC). However, and despite its crucial role, it is still largely unknown how trust emerges, develops, and supports human relationship to technological systems. In the following paper we review the literature on trust, human-robot interaction, HRC, and human interaction at large. Early models of trust suggest that it is a trade-off between benevolence and competence; while studies of human to human interaction emphasize the role of shared behavior and mutual knowledge in the gradual building of trust. We go on to introduce a model of trust as an agent' best explanation for reliable sensory exchange with an extended motor plant or partner. This model is based on the cognitive neuroscience of active inference and suggests that, in the context of HRC, trust can be casted in terms of virtual control over an artificial agent. Interactive feedback is a necessary condition to the extension of the trustor's perception-action cycle. This model has important implications for understanding human-robot interaction and collaboration–as it allows the traditional determinants of human trust, such as the benevolence and competence attributed to the trustee, to be defined in terms of hierarchical active inference, while vulnerability can be described in terms of information exchange and empowerment. Furthermore, this model emphasizes the role of user feedback during HRC and suggests that boredom and surprise may be used in personalized interactions as markers for under and over-reliance on the system. The description of trust as a sense of virtual control offers a crucial step toward grounding human factors in cognitive neuroscience and improving the design of human-centered technology. Furthermore, we examine the role of shared behavior in the genesis of trust, especially in the context of dyadic collaboration, suggesting important consequences for the acceptability and design of human-robot collaborative systems.

Keywords: trust, control, active inference, human-robot interaction, cobotics, extended mind hypothesis, human computer interaction

INTRODUCTION

Technology greatly extends the scope of human control, and allows our species to thrive by engineering (predictable) artificial systems to replace (uncertain) natural events (Pio-Lopez et al., 2016). Navigating and operating within the domain of regularities requires considerably less motor and cognitive effort (e.g., pressing a switch to lift heavy weights) and less perceptual and attentional resources (Brey, 2000); thereby increasing the time and energy available for other activities. However, the inherent complexity of technological systems invariably leads to a state of "epistemic vulnerability," whereby the internal dynamics of the system are hidden to the user and, crucially, must be inferred from the observer via the behavior of the system. Indeed, current misgivings about machine learning rest upon the issue of explainability and interpretability namely, the extent to which a user can understand what is going on "under the hood" (Došilović et al., 2018). By epistemic vulnerability here we mean that the user relies on inference to understand the machine–what the machine does, how it does it, how its actions change given context, etc. Critically, the lack of opacity of these processes may give rise to suspicions and qualms regarding the agent's goals. What factors influence trust during human-robot interaction, and how does human inference modulate the continuous information exchange in human-computer systems? It is widely recognized that trust is a precondition to (successful) human-machine interactions (Lee and See, 2004; Sheridan, 2019a). However, despite great effort from researchers in the field, we still lack a computational understanding of the role of trust in successful human interactions with complex technological systems. Here, we review contemporary theories of trust and their associated empirical data in the context of human-machine interaction. Drawing on the literature in cognitive science of active inference (Friston et al., 2006), control (Sheridan, 2019b), and hierarchical perception-action cycles (Salge and Polani, 2017), we introduce a cross-disciplinary framework of trust–modeled as a sense of *virtual control*. To understand the role of trust in robotics, we first present a brief overview of basic cognitive functions, focusing on the organization of motor control. We then explain the fundamental components of trust–in terms of active inference–and conclude with some remarks about the emergence and development of trust in the context of dyadic human-robot collaboration (HRC), which we take as a good use case for this approach to trust.

SURPRISE MINIMIZING AGENTS

From the standpoint of contemporary cognitive neuroscience, perception and action are means for living organisms to reduce their surprise (i.e., acquire information) about (past, current, and future) states of the world (Friston et al., 2006). The brain according to this framework is considered to be a constructive, statistical organ that continuously generates hypotheses (i.e., beliefs) to predict the most likely causes of the sensory data it encounters (i.e., sensations). These predictions then guide behavior accordingly in a top-down fashion (Gregory, 1980). Various unifying and complementary theories have been proposed to describe this process (e.g., the free energy principle, active inference, predictive processing, dynamic logic, and the Bayesian brain hypothesis). Three fundamental brain functions are defined as follows: (1) perception senses change in the surroundings, (2) cognition predicts the consequences of change, and (3) action controls the causes of change. This tripartition is reflected in the hierarchical functional architecture of brain systems (Kandel et al., 2000), speaking to the brain as an engine of prediction ultimately aiming at the minimization (and active avoidance) of surprising states (see **Figure 1**). There are several ways of describing the requisite (neuronal) message passing–in terms of Bayesian belief updating (Friston et al., 2017). Perhaps the most popular at present is predictive coding (Rao and Ballard, 1999), where inference and learning is driven by prediction errors, and agency emerges from perception-action loops (Fuster, 2004; Parr and Friston, 2019), continuously exchanging information with the sensorium. By sense of agency we refer to the feeling of control over one's actions and their perceived consequences (Gallagher, 2000; Haggard, 2017).

As underwriting perception and action (Méndez et al., 2014), cognition (i.e., active inference or planning) is closely related to evaluating the consequences of action in relation to prior beliefs about homeostatic needs of survival and reproduction; preparing responses to anticipated change (Pessoa, 2010). Here, beliefs correspond to Bayesian beliefs (i.e., posterior probability distributions over some hidden state of the world)–as opposed to propositional beliefs in the folk psychology sense. Minds and their basic functions–such as perception, emotion, cognition, and action–ultimately seek good predictive control. That is, they are continuously aiming to minimize uncertainty about states of the world, where uncertainty is simply expected surprise (i.e., entropy), given a course of action. There are two fundamental ways to avoid (expected) surprise: (1) change one's cognition, beliefs or hypotheses (i.e., perception), or (2) change the world (i.e., action). This distinction is crucial in the context of robotic systems, which are quintessentially concerned with changing the causes of sensations, rather than changing perceptual inference via cognition (Jovanović et al., 2019).

In short, action aims at reducing uncertainty, where exploratory behavior leads us to interact "freely" with objects in the world–to improve our generative models of the way they behave, maximizing the fit between them, and ultimately rendering these behaviors more predictable (Pisula and Siegel, 2005). A generative model is at the heart of active inference–and indeed the current treatment. Technically, models are a probabilistic specification of how (sensory) consequences are caused by hidden or latent states of the world. It generally comprises a likelihood; namely, the probability of a sensory outcome given a hidden state–and prior beliefs over hidden states. Maximizing the fit or alignment between a generative model of the sensed world–and the process generating sensory outcomes corresponds to minimizing surprise (e.g., prediction error) or–in more statistical terms–maximizing the evidence for their model (Hohwy, 2016). In the setting of active inference, this is often referred to as self-evidencing. In active inference, (expected) surprise is approximated with (expected) variational

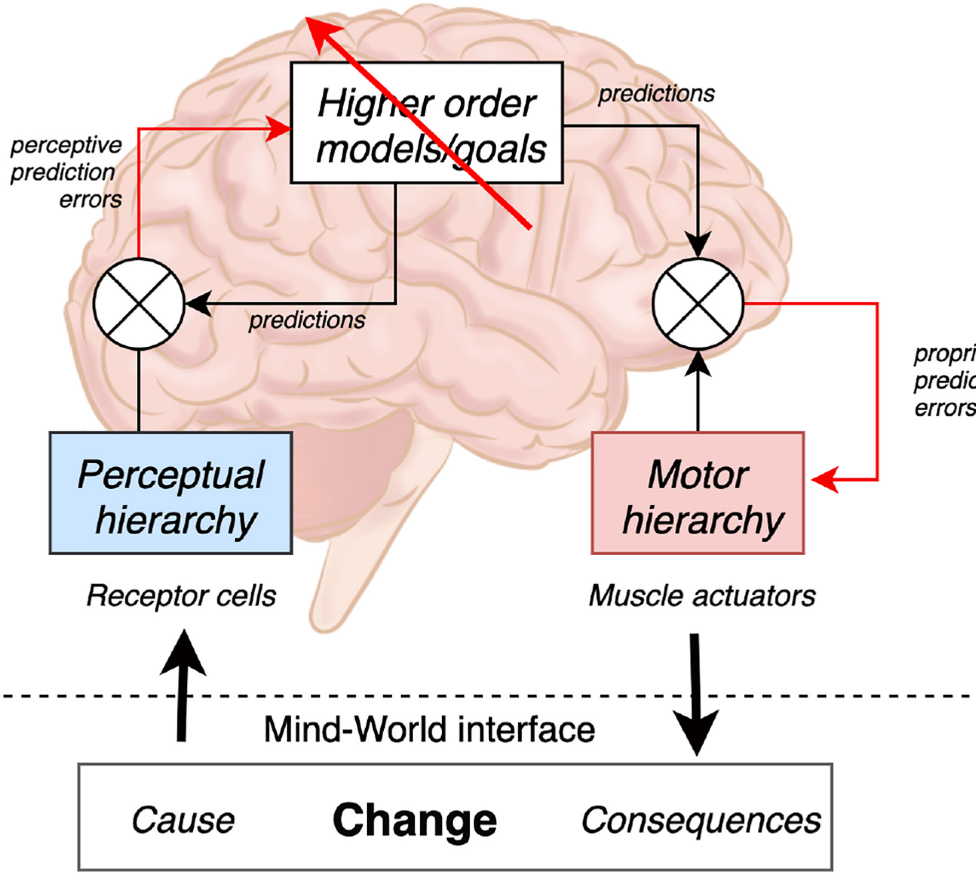

FIGURE 1 | Perception models afferent changes in states of the world detected by receptor cells (e.g., in the retina) all along the perceptual hierarchy. In this control diagram, ⊗ denotes a comparator. The red arrows denote inference and learning (i.e., driven by prediction errors) that compare (descending) predictions with (ascending) sensations. Cognition and higher order processing attempt to predict sensory input and futures states of the world based on available (generative) models; thereby, minimizing prediction error. Action organizes the motor hierarchy in an attempt to actively control the efferent consequences of ongoing events; namely, by modifying causes anticipated through perceptual means, thereby altering the system' dynamics to make them more predictable (i.e., less surprising). Though not specified on this diagram, perception can be further subdivided into interoception and exteroception; respectively, modeling changes in the internal and external world. Emotion–and related notions of selfhood–usually arise via predictive processing of interoceptive sensations, often known as interoceptive inference (Seth, 2013, 2014; Seth and Friston, 2016).

free energy; thereby providing a tractable objective function for perception and action. The integration of efferent (motor) and afferent (sensory) signals results in what can be termed the sensation of control, or feeling of agency, whereby sensorimotor mismatch is minimized.

These three functions of perception-cognition-action form a hierarchical system with sensorimotor signals at the lowest levels of the hierarchy, and abstract cognition (executive functions of goal- directed planning and decision-making) at the highest levels (Schoeller et al., 2018). Perception is organized in a hierarchical fashion, with bottom-up sensory signals (e.g., "a change in color from red to green") being continuously predicted by top-down cognitive models (e.g., "green-light authorization for crossing the street"). Action models are also organized hierarchically, whereby fine motor interaction with the external world (e.g., typing on a keyboard), are contextualized by higher order goals (e.g., writing a paragraph), themselves prescribed by high abstract plans (e.g., getting a paper accepted in a conference)–ultimately underwriting existential goals–corresponding to the organization of life itself (Schoeller et al., 2018).

A key notion is precision weighting, which refers to the reliability or salience of prediction errors. The higher the precision, the more impactful the prediction errors on how processing unfolds. In Active Inference terms, precision represents the agent's confidence that certain action policies (i.e., sequence of actions) will produce the states the agent highly expects (Friston et al., 2014). Predictive agents decide what actions to pursue based on the predicted sensory consequences of the action–choosing those behaviors that are most likely to minimize surprise over the long term, and so maximize their time spent in the sensory states they expect. The performance of action policies to reduce prediction error can be plotted as a slope that depicts the speed at which errors are predicted to be managed along the way. The steepness of the slope indicated how fast errors are being reduced given some policy: the steeper the slope the faster the rate, the shallower the slope the slower the rate. If the speed of error reduction is faster that expected, the action policy should be made more precise; and if the rate is slower than expended, and errors are amassing unexpectedly, then the policy isn't as successful at bringing about those future sensory states that are expected, and this should be taken as evidence for weighing an action policy as having low precision.

Change in the rate at which error is being resolved manifests for humans as emotional valence–we feel good when error is being reduced at a better than expected rate, and we feel bad when error is unexpectedly on the rise (Joffily and Coricelli, 2013; Schoeller, 2015, 2017; Schoeller and Perlovsky, 2016; Schoeller et al., 2017; Van de Cruys, 2017; Kiverstein et al., 2019; Perlovsky and Schoeller, 2019; Wilkinson et al., 2019; Nave et al., 2020). Valence systems provide the agent with a domain general controller capable of tracking changes in error managements and adjusting precision expectations relative to those changes (Kiverstein et al., 2019; Hesp et al., 2021). This bodily information is a reflection of an agent's perceived fitness–that is, how adaptive the agent's current predictive model is relative to their environment.

Affective valence is widely acknowledged to play an important role in trust (Dunn and Schweitzer, 2005). Positive feelings have been shown to increase trusting, while negative feelings diminish it (Dunn and Schweitzer, 2005). The active inference framework helps to account for this evidence, suggesting that positive and negative feelings are in part a reflection of how well or poorly one is able to predict the actions of another person. As detailed in the following section, affectivity plays a crucial role in mediating exchanges with robots, often acting as a cardinal determinant of trust in that context specifically (Broadbent et al., 2007). As a consequence, robotic design that considers affect–and related higher-level constructs–are likely to enhance productivity and acceptance (Norman et al., 2003).

AGENCY AND EMPOWERMENT IN HUMAN-TECHNOLOGICAL EXTENSION

The relevance of active inference for robotics has been experimentally demonstrated in Pio-Lopez et al. (2016). In the context of automation, understanding human agency is all the more important–as experimental studies have demonstrated that one can prime for agency with external cues (leading to abusive control), and clinical studies reveal that an impairment of control is associated with depression, stress, and anxiety-related disorders (Abramson et al., 1989; Chorpita and Barlow, 1998). The integration of efferent (motor) and afferent (sensory) signals results in what can be termed the sensation of control or a feeling of agency (Salomon et al., 2016; Vuorre and Metcalfe, 2016), which depends on the correspondence of top-down (virtual) predictions of the outcomes of action, and the bottom-up (actual) sensations. As illustrated in **Figure 1**, the brain compares actual sensory consequences of the motor action with an internal model of its predicted sensory consequences. When predicted sensory consequences match incoming sensory signals, the movement is attributed to the self and a (confident) sense of agency is said to emerge (Wolpert et al., 1995; Hohwy, 2007; Synofzik et al., 2008; Salomon et al., 2016). Situations where there is a mismatch between intended and observed actions we also see a feeling of loss of agency, and an attribution of the movement (or lack thereof) to an external source. For example, if someone was to move my arm then there would be the sensory experience but without the prediction. If instead I was to try to move my arm, but due to anesthetic I was unable to, there would be the prediction but not the sensory confirmation. Agency then is just another hypothesis (or Bayesian belief) that is used to explain interoceptive, exteroceptive, and proprioceptive input. If sensory evidence is consistent with my motor plans, then I can be confident that "I caused that." Conversely, if I sense something that I did not predict, then the alternative hypothesis that "you caused that" becomes the best explanation (Seth, 2015). The accompanying uncertainty may be associated with negative affect such as stress or anxiety (Stephan et al., 2016; Peters et al., 2017). Again, the very notions of stress and anxiety are treated as higher-level constructs–that best explain the interoceptive signals that attend situations of uncertainty and adjust precision accordingly; e.g., physiological autonomic responses of the flight or fright sort (Barrett and Simmons, 2015; Seth and Friston, 2016).

To measure the amount of control (or influence) an agent has and perceives, Klyubin et al. (2005) proposed the concept of empowerment. Empowerment is a property of self-organized adaptive systems and is a function of the agent perception-action loop, more specifically the relation between sensors and actuators of the organism, as induced by interactions between the environment and the agent's morphology (Salge and Polani, 2017). Empowerment is low when the agent has no control over what it senses, and it is high the more control is evinced (Friston et al., 2006). An information-theoretic definition has been proposed, whereby empowerment is interpreted as the amount of information the agent can exchange with its environment through its perception-action cycle. According to Klyubin et al. (2005), empowerment is null when the agent has no control over what it is sensing, and it is higher the more perceivable control or influence the agent has. Hence, "empowerment can be interpreted as the amount of information the agent could potentially inject into the environment via its actuator and later capture via its sensor." Consider for example the difference between passively watching a movie and being engaged with the same content in an immersive virtual reality setting. Crucially, empowerment is a reflection of what an agent *can* do, not what the agent actually does (Klyubin et al., 2005), and maximizing empowerment adapts sensors and actuators to each other. In other words, empowerment can be described in terms of sensorimotor fitness–i.e., the spatial and temporal relevance of the feedback the robot gets on its own behavior. For example, a robot that gets multisensor feedback on the probability of success of its actions has greater empowerment than a robot who is deprived of, say, visual information or which receives delayed information (the greater the delay, the weaker the empowerment). This calls forth a framework where the so-called exploration/exploitation dilemma (crucial for safety in HRC) can be casted as a behavioral account of the perception-action cycle.

Technology considerably increases human empowerment (Brey, 2000), freeing the human animal from many niches or geographical constraints (e.g., climate or geology), and allowing increasingly complicated narratives and trajectories to develop within the scope of human control (e.g., cranes allow the manipulation of heavy systems beyond mere human capabilities). Predictive organisms are attracted to–and rewarded by–opportunities to improve their predictive grip on their environments–i.e., to improve their empowerment. By definition, technological extension of the perception-action cycle offers a powerful way of expanding empowerment, but to function effectively it needs to be integrated with the agent's sensorimotor dynamics. In other words, technology must enter the agent' extended repertoire of behaviors. That inclusion requires the technological extensions to be modeled internally by the agent in the same capacity of its own sensorimotor contingencies, at some level of abstraction. This (self) modeling of technological extension is key to the emergence of trust–in active inference terms: a high precision on beliefs about how the technology will behave and evolve relative to our own sensorimotor engagements. This is an extension of the same mechanism giving rise to agency beyond the realm of the body. As we attempt to show in the next section, this extension of human control beyond mere motor action and its cognitive monitoring requires trust–as a sense of virtual control in an extended perception-action cycle (Sheridan, 1988). The study of human agency has clear relevance for robotic motor control, but to our knowledge it has not yet been applied to the problem of trust in complex technological systems or human-robot interaction. In the next section, we examine the possibility of modeling trust in relation to active inference and empowerment.

TRUST AS VIRTUAL CONTROL IN EXTENDED AGENCY

Within the context of human-robot interactions (Lee, 2008), optimal trust is crucial to avoid so-called disuse of technology (i.e., loss of productivity resulting from users not trusting the system), but also abuse of technology (i.e., loss of safety resulting from overreliance on the system). Hence, the cognitive neuroscience of trust has implications for both safety and management (Sheridan and Parasuraman, 2005; Lee, 2008). Indeed, technological abuse and overreliance on automation count among the most important sources of catastrophes (Sheridan and Parasuraman, 2005). From a theoretical point of view, tremendous variations exist in what trust represents and how it can best be quantified, and several definitions have been suggested with potential applications for automation (Muir, 1994; Cohen et al., 1999). An exhaustive review–of the large body of work devoted to trust literature–is outside the scope of this article: excellent reviews can be found in Lee and See (2004) and Sheridan (2019b). Here, we present the fundamental elements of these models of trust, in the light of perception-action loops, and potential applications to robotics to demonstrate the relevance of the active inference framework for human factors in HRI.

Several measures of trust exist in a variety of settings from management, to interpersonal, and automation. In reviewing the literature on trust, Lee and See identified three categories of definitions; all fundamentally related to uncertainty and control (2004). The fundamental relation between trust and uncertainty appears most salient in situations when the uncertainty derives from the realization of goals or intentions (e.g., in human-robot interactions, or employee-employer relationships), where internal details about the agent are unknown, leaving the trustor vulnerable. In the context of robotics–where human action is extended by robotic systems–the match between goals of the (extended) human agent and those of the (extending) robotic agent is crucial in determining the success of the relation (whether the agent will make use of the extension). In order of generality, the definitions identified by Lee and See are: (1) trust as intention to (contract) vulnerability, (2) trust as vulnerability, and (3) trust as estimation of an event likelihood. Note that these three general definitions, derive from early definitions of trust by Muir (1994) and Mayer et al. (1995), according to whom trust is a trade-off between ability (A) and benevolence (B), whereby a reliable system is high in both A and B (**Figure 2**).

The importance of externalizing goals of robotics systems (i.e., transparency) at all levels of the hierarchical perception-action loop cannot be stressed enough–for successful communication and gradual building of trust (Sheridan and Parasuraman, 2005). This is well captured in the standard definition of trust by Sheridan (2019b), where communication of goals (or transparency) plays a crucial role among the seven item scales of trust (see **Table 1**).

In summary, trust is fundamentally related to human control to the extent that it is required for any extension of the perception-action cycle (i.e., when the success of the performance depends on some other agent's perception-action cycle, rather than one's own). Above, we saw that vulnerability is a function of empowerment in the extended agent (the more extended the agent, the more vulnerable), which can be evaluated through interaction with the robotic perception-action cycle. This may help to explain why operator curiosity is an important source of accidents in the robot industry (Lind, 2009), as curiosity aims to reduce uncertainty about the technology and so increase trust and control, and suggests potential solutions in the field of

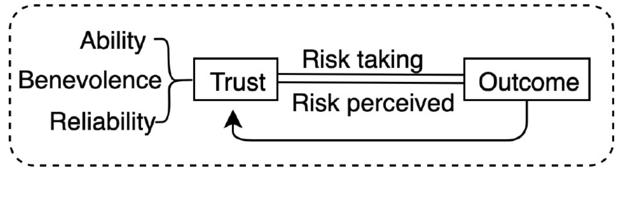

FIGURE 2 | Muir and Mayer model of trust as a function of the trustee's ability, benevolence and reliability (1995) where risk perception affects risk action. This bipartition of trust as ability and benevolence amounts to two different levels in the motor hierarchy of the extended agent (e.g., the robot), whereby benevolence refers to the high-level goals motivating the extended agent and ability refers to the means of the agent to realize these goals, i.e., the sophistication of its low-level motor output in relation to the task at hand.

TABLE 1 | Standard definition of trust by Sheridan (2019b).

(1) Statistical reliability (lack of error).

(2) Usefulness (ability of the system to do what is most important, e.g., in trading benefits and costs).

(3) Robustness (ability and flexibility of the system to perform variations of the task).

(4) Understandability (transparency of the system in revealing how and why it is doing what it is doing).

(5) Explication of intent (system communicating to the trustee what it will do next).

(6) Familiarity (to the user based on past experience).

(7) Dependence (upon the system by the trustee as compared to other ways of doing the given task).

accidentology. Trust is required in situations of uncertainty; and it varies as the system exhibits predictable regularities. Sheridan and Meyer models suggest that one will trust a predictable system, to the extent that one can act upon that system to obtain similar results over time, and eventually render its behavior more predictable through incremental alterations.

We have considered how a sense of agency emerges, as the resolution of mismatch between (1) the (perceptual) expectation (i.e., hypothesis) about the consequences of (motor) action, and (2) the perceived results of action (observation, perception). We introduced the idea of trust as a sense of virtual, extended control. In other words, trust is a measure of the precision, or confidence, afforded by action plans that involve another (i.e., of the match between one's actions–and their underlying intentions–and the predicted sensory consequences through another agent). As such, "trust" is an essential inference about states of affairs; in which the anticipated consequences of extended action are realized reliably. From the point of view of "emotional" inference (Smith et al., 2019), trust is therefore the best explanation for a reliable sensory exchange with an extended motor plant or partner. Given the role that affect plays in tuning precision on action policies, "reliable" here means a reliable way to reduce expected free energy (via the extended interaction). We are attracted by, or solicited to use, a tool or device because it affords to us a means of reducing error, in a better than expected way relative to doing the same work in the absence of technological extension.

It is generally assumed that trust in any system increases with evidence of that system's reliability (**Figure 3**). The greater the convergence of behavior models between trustor and trustee (i.e., the largest the benevolence), the greater the trust in the relationship (Hisnanick, 1989). Perhaps, this explains why simple mimicry facilitates adoption, or why one tends to agree with people who behave similarly–we generalize shared goals on the basis of shared behavior (Cirelli, 2018). The similarity-attraction hypothesis in social psychology predicts that people with similar personality characteristics will be attracted to each other (Morry, 2005). Hence, technology that displays personality characteristics–similar to those of the user–tends to be accepted more rapidly (Nass et al., 1995). As machines become increasingly intelligent, it is crucial that they communicate higher-order goals accordingly (Sheridan, 2019b). Communication of goals can be simplified by rendering the perception-action cycle explicit/and augmenting sensors to indicate their perceptual range (e.g., the human retina affords some information about the portion of the visual field it senses); thereby, greatly reducing the risk of accidents.

Finally, trust is a fundamentally dynamic process that eventually leads to a state of dependence (**Figure 4**). This is best exemplified in the context of information technology, whereby the information is no longer stored internally (e.g., phone numbers, navigation pathways, historical facts) but all that is known is the access pathway (my phone's contact list, my preferred web mapping service, a Wikipedia page). As suggested by the Sheridan scale, the dynamics of trust go beyond mere predictability and ultimately lead to a state of prosthetic dependence in the context of the specific task. This is evident in the context of automation, which increases the perception-action cycle at an exponential rate, thereby leading to a high abandon rate of past practices, as new technologies are adopted. Formally speaking, as technology allows the agent to reduce prediction error (by better understanding the problem space, and through more empowered actions) the agent comes to expect that slope of error reduction within those contexts and relative to the specific tasks. The result is a gradual loss of interest or solicitation by previous less potent forms of HRCs–they have become outdated and so have lost their motivational appeal.

In the context of interpersonal relationships, Rempel et al. (1985) described trust as an evolving phenomenon, where growth is a function of the relationships progress. They further argue that the anticipation of future behavior forms the basis of trust at the earliest stages of a relationship. This is followed by dependability, which reflects the degree to which behavior is consistent. As the relationship matures, the basis of trust ultimately passes the threshold of faith, which has been related to benevolence (Lee and See, 2004); i.e., coordination on higher order goals driving behavior. Crucially, an early study of the adaptation of operators to new technology demonstrated a similar progression (Hisnanick, 1989). Trust in that context depends on trial-and-error experience, followed by understanding of the technology's operation, and finally, a state of certainty or faith (see **Figure 5**). Lee and Moray (1992) made similar distinctions in defining the factors that influence trust in automation.

TRUST DURING DYADIC COLLABORATION

We have seen that the essential components of trust (benevolence and competence) can be cast in terms of the confidence in beliefs at (respectively) high and low levels in the motor hierarchy, but how can active inference contribute to the science of extended agency? In this section, we examine the role of

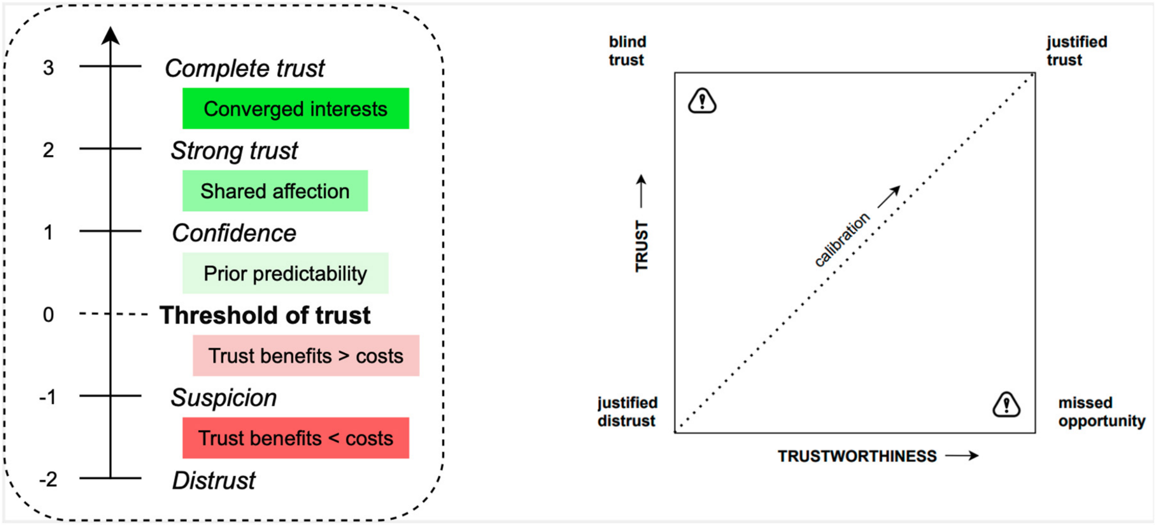

FIGURE 3 | On the left, levels of trust from Dietz and Den Hartog (2006). On the right, cross-plot of (objective) trustworthiness compared to (subjective) trust by Sheridan (1988), Sheridan (2019b). As a pioneer in the study of trust in technology, Sheridan further suggested that (subjective) trust can be cross-plotted against (objective) trustworthiness. This representation engenders four extremes: justified trust or distrust, blind trust (trusted untrustworthy; i.e., misuse) and missed opportunity (untrusted trustworthy; i.e., disuse). The dotted curve represents calibration, which is linear when trust is justified. Poor calibration can lead to loss of safety (due to overconfident misuse), or loss of productivity (due to underconfident disuse).

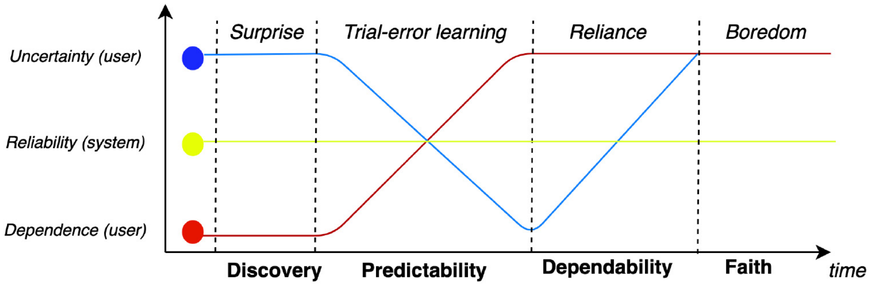

FIGURE 4 | Dynamics of trust over time–with four phases from discovery to faith: for a consistently reliable system, dependence (i.e., risk) is inversely proportional to uncertainty, assessed through a cycle of trial and error, until a threshold is reached. Through cycles of trial and errors, trust gradually evolves from predictability (model) to dependability (control) to a state of faith (overreliance). Our model suggests that boredom is a marker of overreliance.

expectations in the context of dyadic interaction. So, what would a formal (first principles) approach like active inference bring to HRC? At its most straightforward, trust is a measure of the confidence that we place in something behaving in beneficial ways that we can highly predict. Technically, this speaks to the encoding of uncertainty in generative models of dyadic interactions. These generative models necessarily entail making inferences about policies; namely, ordered sequences of action during dyadic exchanges (Moutoussis et al., 2014; Friston and Frith, 2015). This could range from turn taking in communication (Wilson and Wilson, 2005; Ghazanfar and Takahashi, 2014) to skilled interactions with robotic devices. At its most elemental, the encoding of uncertainty in generative models is usually framed in terms of the precision (i.e., inverse variance) or confidence (Friston et al., 2014). Crucially, every (subpersonal) belief that is updated during active inference can have the attribute of a precision or confidence. This means that the questions about trust reduce to identifying what kind of belief structure has a precision that can be associated with the construct of "trust." In generative models based upon discrete-state spaces (e.g., partially observed Markov decision processes) there are several candidates for such beliefs. Perhaps the most pertinent–to dyadic interactions–are the beliefs about state transitions; i.e., what happens if I (or you) do that. For example, if I trust you, that means I have precise Bayesian beliefs about how you will respond to my actions. This translates into precise beliefs about state transitions during controlled exchanges (Parr and Friston, 2017; Parr et al., 2018). This means that I can plan deep into the future before things become very uncertain and, in turn, form precise posterior beliefs about the best courses of action, in other words our policies align (see **Figure 5**).

Conversely, if I do not trust you, I will have imprecise beliefs about how you will respond and will only be able to entertain

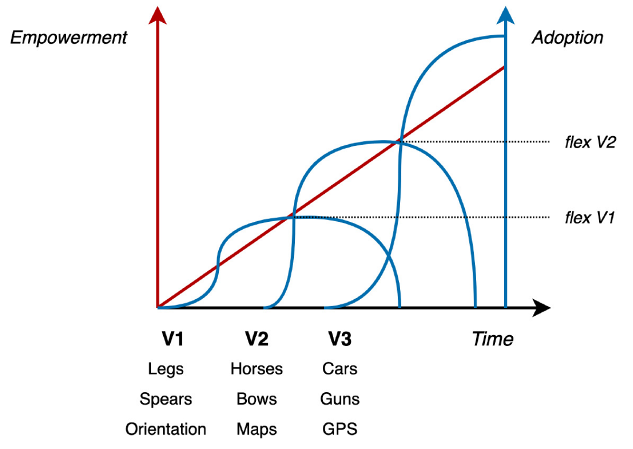

FIGURE 5 | Waves of technological adoption related to predictive slopes of extended engagement (empowerment) during versioning of the technology. Indeed this is an oversimplification for the sake of visualization as we are assuming a linear progression of empowerment over time in the evolving versions of the technology (i.e., a healthy research and development cycle) where, for most technologies, newer versions may not present much greater empowerment as compared to older ones. The important idea here is the inflection point (flex) indicating the start of technological decay reflecting the abandon rate of a practice as the experience of better predictive slopes of extended technological engagements lead to disengagement of non-extended approaches (e.g., cars replace horses replacing legs). Old slopes are less than expected and so unsatisfactory as compared to new ones.

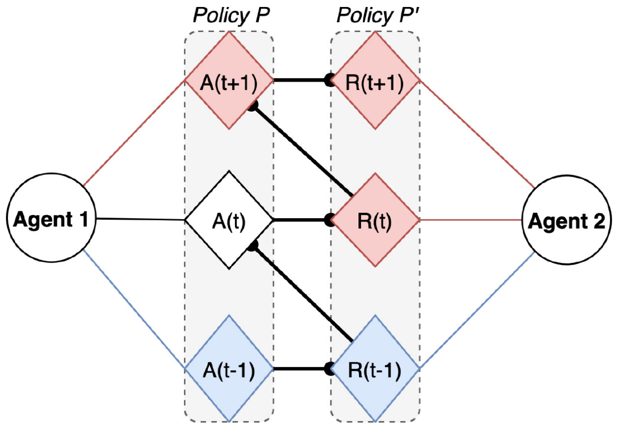

FIGURE 6 | A trust dyad, whereby Agent 1 performs action A at t, within the action policy P, and in collaboration with Agent 2. In a trustworthy relationship, Agent 1 can expect from Agent 2 an action Policy P', where P' is symmetrical to P (each action of P' at t + 1 is a response to P at t). Past (observed) actions are blue and future (anticipated) actions are red. The bold line in between policies represents the shared policy (or joint narrative), whereby A(t + 1) can be prepared based on beliefs about anticipated R(t).

short term plans during any exchange. Furthermore, it will be difficult to infer precise outcomes of any course of action–and hence hard to entertain a shared policy. This means I will also be uncertain about which is the best course of action. Technically, this results in an imprecise belief distribution over policies, which is normally associated with negative affect or some form of angst (Seth and Friston, 2016; Badcock et al., 2017; Peters et al., 2017). Notice, that now there is not just error in the environment to deal with but also the uncertainty of the other. As uncertainty increases, negatively valenced feelings emerge as a reflection of that change, and in turn reduce precision on the policies related to that collaboration. The result is the agent is less likely to be attracted to enact policies of extension with that other person or robot, and so much more likely to revert to using more habitual (and already highly precise) ways of reducing error. In short, almost by definition, engaging with an untrustworthy partner is, in a folk psychological sense, rather stressful.

Clearly, this active inference formulation is somewhat hypothetical. There will be many other belief structures that could be imprecise; for example, prior beliefs about the policies I should entertain and, indeed, the precision of likelihood mappings (that map from latent or hidden states of the world to observed outcomes). The latter is usually considered in terms of ambiguity (Friston et al., 2017; Veissière et al., 2019). In other words, I could consider your behavior or responses ambiguous–and that could render you untrustworthy; even if I have very precise beliefs about the latent states you are likely to navigate or pursue. In short, it may be an open question as to whether the precision of state transitions, likelihood contingencies or prior beliefs about policies manifest as differences in trust. This brings us to a fundamental motivation for the formalization of trust in terms of active inference.

It is possible to build models of dyadic exchange under ideal Bayesian assumptions using active inference (e.g., Moutoussis et al., 2014; Friston and Frith, 2015). This means that one can optimize the prior beliefs inherent in these models to render observed choice behavior the most likely. Put another way, one can fit active inference models to empirical behavior to estimate the prior beliefs that different subjects evince through their responses (Parr et al., 2018). These estimates include a subject's prior beliefs about the precision of various probability distributions or Bayesian beliefs. In turn, this means it should be possible to phenotype any given person in an experimentally controlled (dyadic) situation and estimate the precision of various beliefs that best explain their behavior. One could, in principle, then establish correlations between different kinds of precision and other validated measures of trust, such as those above. This would then establish what part of active inference best corresponds to the folk psychological–and formal definitions of trust. Interestingly, this kind of approach has already been considered in the context of computational psychiatry and computational phenotyping; especially in relation to epistemic trust (Fonagy and Allison, 2014). Epistemic trust is a characteristic of the confidence placed in someone as a source of knowledge or guidance. Clearly, this kind of trust becomes essential in terms of therapeutic relationships and, perhaps, teacher pupil relationships. Finally, one important determinant of the confidence placed in–or precision afforded–generative models of interpersonal exchange is the degree to which I can use myself as a model of you. This speaks to the fundamental importance of a shared narrative (or generative model) that underwrites any meaningful interaction of the sort we are talking about. This can be articulated in terms of a generalized synchrony that enables a primitive form of communication or hermeneutics (Friston and Frith, 2015). Crucially, two agents adopting the same model can predict each other's behavior, and minimize their mutual prediction errors (**Figure 6**). This has important experimental implications, especially in the context of HRC, where robotic mimicry can be seen as mere self-extension for the user, leading to what philosophers of technology call relative transparency (where whatever impacts the robot also impacts me–see Brey, 2000). The self being the product of the highest prediction capacities, when another agent becomes more predictable it also increases the similarity at the highest levels in the cognitive hierarchy and thereby facilitates joint action.

This mutual predictability is also self-evident in terms of sharing the same narrative; e.g., language. In other words, my modeling of you is licensed as precise or trustworthy if, and only if, we speak the same language. This perspective can be unpacked in many directions; for example, in terms of niche construction and communication among multiple conspecifics (in an ecological context) (Constant et al., 2019; Veissière et al., 2019). It also speaks to the potential importance of taking into account self-models in HRC design, allowing both users and robots to represent each other's behavior efficiently. Indeed, on the above reading of active inference, such shared narratives become imperative for trustworthy exchanges and collaboration. Indeed, current models suggest that the rise of subjectivity and the "self" are grounded in privileged predictive capacities regarding the states of the organism compared to the external environment (Limanowski and Blankenburg, 2013; Apps and Tsakiris, 2014; Allen and Friston, 2016; Salomon, 2017). As such, dyadic trust in another agent (biological or artificial) can be viewed as a process of extending these predictive processes beyond the body and rendering the external agent as part of a self model. Moreover, recently robotic interfaces have been used to induce modulations of self models by interfering with sensorimotor predictions. This in turn gives rise to phenomena closely resembling psychiatric symptoms (Blanke et al., 2014; Faivre et al., 2020; Salomon et al., 2020).

CONCLUSION

In the light of our increasing dependence on technology, it is worth considering that the largest aspect of human interactions with machines (their use) essentially rests upon vague approximative mental models of the underlying mechanisms (e.g., few smartphone users can understand the functioning of a computer operating software). Technically, in active inference, the use of simplified generative models (e.g., heuristics) is an integral part of self-evidencing. This follows because the evidence for a generative model (e.g., of how a smartphone works) can be expressed as accuracy minus complexity. In this setting, complexity is the divergence between posterior and prior beliefs–before and after belief updating. This means the generative model is required to provide an accurate account of sensory exchanges (with a smartphone) that is as simple as possible (Maisto et al., 2015). In short, the best generative model will be, necessarily, simpler than the thing it is modeling. This principle holds true of technology in general (extending the scope of human perception-action cycles), and automation specifically (replacing these perception-action capabilities). We have examined the concept of trust from the standpoint of control and perception-action loops and found that trust components (i.e., competence and benevolence) are best casted in terms of an action-cognitive hierarchy. By examining trust from the standpoint of active inference, we were also better able to understand phenomena, such as exploration-related accidents, and the gradual building of trust with shared goals, narratives and agency. One of the benefits of this model is that it applies to any sort of collaborative enterprise between humans and machines. Although the specifications of the machine (e.g., its size, its use, etc.) and the nature of the collaboration (e.g., occasional, constant, autonomous, etc.) will of course change how and what one models about the collaborative machine, the trust one feel emerges from the identical process of modeling their states and behaviors over time in ways that allow them to be included in one's own generative model (in a particular context). HRC is of course only a first step and it will be interesting going forward to consider how this model of trust as extended predictive control practically is applied to the wide variety of cases where humans and machines are working closely together in our world today.

As the complexity and autonomy of artificial systems go up, so too will the complexity and sophistication of the model we generate about the behaviors of those systems. In the case of collaborating with artificial intelligence systems this becomes even more challenging, and would increasingly require useful opacifications of the underlying decision making mechanisms that drive those system's behaviors. The science of human-robot interaction could make rapid progress if objective measures of trust were developed, and the neuroscience of agency does offer such metrics. It is here that a simulation setup of the sort offered by active inference could play an important part. Among the potential biomarkers for agency and control, the N1 component of event related electrical brain responses–a negative potential occurring approximately 100 ms after stimulus onset–is attenuated during self-produced or predicted events, relative to that observed during externally generated feedback. As machine become increasingly intelligent, it is to be expected that not only users will develop more sophisticated (generative) models of their internal behavior and the reliability of these behavior, but robots will also adapt to interindividual differences (Sheridan, 2019b), hence reciprocally monitor the trustworthiness of users, and thereby allow for safer and more productive interaction.

In this paper we have proposed a novel view of trust as extended (predictive) control, a view that is well poised to help us elucidate the mechanisms underlying trust between humans, and between humans and technological artifacts. However, this should only be seen as the beginning. The field of HRC is quickly evolving, as the robots we find ourselves collaborating with are increasingly complex and autonomous. Degree of autonomy is of particular importance here for thinking about HRCs. As autonomy increases in our robotic partners different forms of collaboration are bound to emerge, and new requirements for trusting those artifacts will be necessary. While we do not have the space here to fully explore these more complex examples in current and future HRC, we can at least say that transparency and ethical-design will become increasingly important. Given the framework we have proposed, for trust to emerge in these complex interactions human agents need to be able to accurately (or at least usefully) model the sorts of decision-trees that the autonomous artificial agents make use of in various contexts. The means by which such transparency can be achieved is a topic for further research.

Furthermore, as artificial intelligence systems evolve in complexity we will inevitably be interacting with technological artifacts that are able to model humans in return. This two-way predictive modeling will result in new forms of collaboration and new approaches to developing a trusting relationship (see Demekas et al., 2020). Collaborative dynamics between humans is already being modeled using the AIF (Ramstead et al., 2020), in which predictive agents model each other's generative model in ways that allow groups to temporarily become a unified error-minimizing machine. With the possibility of future artificial autonomous agents using variations of a prediction hierarchy like humans use, exploring the emergent dynamics between human and artificial agents in this way becomes possible as well.

KEY POINTS:

- Mind–all brain–is a constructive, statistical organ that continuously generates hypotheses to predict the most likely causes of its sensory data.
- We present a model of trust as the best explanation for a reliable sensory exchange with an extended motor plant or partner.
- User boredom may be a marker of overreliance.
- Shared narratives, mutual predictability, and self-models are crucial in human-robot interaction design and imperative for trustworthy exchanges and collaboration.
- Generalized synchrony enables a primitive form of communication.
- Shared generative models may allow agents to predict each other more accurately and minimize their prediction errors or surprise, leading to more efficient HRC.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

ACKNOWLEDGMENTS

MM carried out this work with the support of Horizon 2020 European Union ERC Advanced Grant XSPECT – DLV-692739.

REFERENCES

Abramson, L. Y., Metalsky, G. I., and Alloy, L. B. (1989). Hopelessness depression: a theory-based subtype of depression. *Psychol. Rev.* 96, 358–372. doi: 10.1037/0033-295x.96.2.358

Allen, M., and Friston, K. J. (2016). From cognitivism to autopoiesis: towards a computational framework for the embodied mind. *Synthese* 195, 2459–2482. doi: 10.1007/s11229-016-1288-5

Apps, M. A., and Tsakiris, M. (2014). The free-energy self: a predictive coding account of self-recognition. *Neurosci. Biobehav. Rev.* 41, 85–97. doi: 10.1016/j.neubiorev.2013.01.029

Badcock, P. B., Davey, C. G., Whittle, S., Allen, N. B., and Friston, K. J. (2017). The depressed brain: an evolutionary systems theory. *Trends Cogn. Sci.* 21, 182–194. doi: 10.1016/j.tics.2017.01.005

Barrett, L. F., and Simmons, W. K. (2015). Interoceptive predictions in the brain. *Nat. Rev. Neurosci.* 16, 419–429. doi: 10.1038/nrn3950

Blanke, O., Pozeg, P., Hara, M., Heydrich, L., Serino, A., Yamamoto, A., et al. (2014). Neurological and robot-controlled induction of an apparition. *Curr. Biol.* 24, 2681–2686. doi: 10.1016/j.cub.2014.09.049

Brey, P. (2000). "Technology as extension of human faculties," in *Metaphysics, Epistemology, and Technology. Research in Philosophy and Technology*, ed. C. Mitcham (London: Elsevier/JAI Press).

Broadbent, E., MacDonald, B., Jago, L., Juergens, M., and Mazharullah, O. (2007). "Human reactions to good and bad robots," in *Proceedings of the 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems. Presented at the 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems*, (Piscataway, NJ: IEEE).

Chorpita, B. F., and Barlow, D. H. (1998). The development of anxiety: the role of control in the early environment. *Psychol. Bull.* 124, 3–21. doi: 10.1037/0033-2909.124.1.3

Cirelli, L. K. (2018). How interpersonal synchrony facilitates early prosocial behavior. *Curr. Opin. Psychol.* 20, 35–39. doi: 10.1016/j.copsyc.2017.08.009

Cohen, M. S., Parasuraman, R., and Freeman, J. T. (1999). *Trust in Decision Aids: a Model and its Training Implications*. Arlington, VA: Cognitive Technologies. Technical Report USAATCOM TR 97-D-4.

Constant, A., Ramstead, M. J., Veissière, S. P., and Friston, K. (2019). Regimes of expectations: an active inference model of social conformity and human decision making. *Front. Psychol.* 10:679. doi: 10.3389/fpsyg.2019.00679

Demekas, D., Parr, T., and Friston, K. J. (2020). An investigation of the free energy principle for emotion recognition. *Front. Comp. Neurosci.* 14:30. doi: 10.3389/fncom.2020.00030

Dietz, G., and Den Hartog, D. N. (2006). Measuring trust inside organisations. *Personnel Rev.* 35, 557–588. doi: 10.1108/00483480610682299

Došilović, F. K., Brčić, M., and Hlupić, N. (2018). "Explainable artificial intelligence: a survey," in *Proceedings of the 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, (Piscataway, NJ: IEEE), 0210–0215.

Dunn, J. R., and Schweitzer, M. E. (2005). Feeling and believing: the influence of emotion on trust. *J. Pers. Soc. Psychol.* 88, 736–748. doi: 10.1037/0022-3514.88.5.736

Faivre, N., Vuillaume, L., Bernasconi, F., Salomon, R., Blanke, O., and Cleeremans, A. (2020). Sensorimotor conflicts alter metacognitive and action monitoring. *Cortex* 124, 224–234. doi: 10.1016/j.cortex.2019.12.001

Fonagy, P., and Allison, E. (2014). The role of mentalizing and epistemic trust in the therapeutic relationship. *Psychotherapy* 51:372. doi: 10.1037/a0036505

Friston, K., and Frith, C. (2015). A duet for one. *Conscious. Cogn.* 36, 390–405.

Friston, K., Kilner, J., and Harrison, L. (2006). A free energy principle for the brain (archive). *J. Physiol. Paris* 100, 70–87.

Friston, K. J., Parr, T., and de Vries, B. (2017). The graphical brain: belief propagation and active inference. *Network Neurosci. (Cambridge, Mass)* 1, 381–414. doi: 10.1162/netn_a_00018

Friston, K. J., Stephan, K. E., Montague, R., and Dolan, R. J. (2014). Computational psychiatry: the brain as a phantastic organ. *Lancet Psychiatry* 1, 148–158. doi: 10.1016/S2215-0366(14)70275-70275

Fuster, J. M. (2004). Upper processing stages of the perception-action cycle. *Trends Cogn. Sci.* 8, 143–145. doi: 10.1016/j.tics.2004.02.004

Gallagher, S. (2000). Philosophical conceptions of the self: implications for cognitive science. *Trends Cogn. Sci.* 4, 14–21. doi: 10.1016/s1364-6613(99)01417-5

Ghazanfar, A. A., and Takahashi, D. Y. (2014). The evolution of speech: vision, rhythm, cooperation. *Trends Cogn. Sci.* 18, 543–553. doi: 10.1016/j.tics.2014.06.004

Gregory, R. L. (1980). Perceptions as hypotheses (archive). *Phil. Trans. R. Soc. Lond. B* 290, 181–197.

Haggard, P. (2017). Sense of agency in the human brain. *Nat. Rev. Neurosci.* 18, 197–208.

Hesp, C., Smith, R., Parr, T., Allen, M., Friston, K. J., and Ramstead, M. J. (2021). Deeply felt affect: the emergence of valence in deep active inference. *Neural Comput.* 33, 398–446. doi: 10.1162/neco_a_01341

Hisnanick, J. (1989). In the age of the smart machine: the future of work and power. *Emp. Respons. Rights J.* 2, 313–314.

Hohwy, J. (2007). The sense of self in the phenomenology of agency and perception. *Psyche* 13, 1–20.

Hohwy, J. (2016). The self-evidencing brain. *Noûs* 50, 259–285. doi: 10.1111/nous.12062

Joffily, M., and Coricelli, G. (2013). Emotional valence and the free-energy principle. *PLoS Comp. Biol.* 9:e1003094. doi: 10.1371/journal.pcbi.1003094

Jovanović, K., Petrič, T., Tsuji, T., and Oddo, C. M. (2019). Editorial: human-like advances in robotics: motion, actuation, sensing, cognition and control. *Front. Neurorobot.* 13:85. doi: 10.3389/fnbot.2019.00085

Kandel, E., Schwartz, J., and Jessell, T. (2000). *Principles of Neural Science*, 4th Edn. New York City, NY: McGraw Hill Companies.

Kiverstein, J., Miller, M., and Rietveld, E. (2019). The feeling of grip: novelty, error dynamics, and the predictive brain. *Synthese* 196, 2847–2869. doi: 10.1007/s11229-017-1583-9

Klyubin, A. S., Polani, D., and Nehaniv, C. L. (2005). "Empowerment: a universal agent-centric measure of control," in *Proceedings of the Congress on Evolutionary Computation*, (Piscataway, NJ: IEEE), 128–135.

Lee, J., and Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics* 35, 1243–1270. doi: 10.1080/00140139208967392

Lee, J., and See, K. (2004). Trust in automation: designing for appropriate reliance. *Hum. Factors* 46, 50–80. doi: 10.1518/hfes.46.1.50.30392

Lee, J. D. (2008). Review of a pivotal human factors article: "humans and automation: use, misuse, disuse, abuse.". *Hum. Factors J. Hum. Factors Ergonom. Soc.* 50, 404–410. doi: 10.1518/001872008x288547

Limanowski, J., and Blankenburg, F. (2013). Minimal self-models and the free energy principle. *Front. Hum. Neurosci.* 7:547. doi: 10.3389/fnhum.2013.00547

Lind, S. (2009). *Accident Sources in Industrial Maintenance Operations. Proposals for Identification, Modelling and Management of Accident Risks (Tapaturmat teollisuuden kunnossapitotöissä - Ehdotuksia tapaturmariskien tunnistamiseen, mallinnukseen ja hallintaan)*. Espoo: VTT Publications.

Maisto, D., Donnarumma, F., and Pezzulo, G. (2015). Divide et impera: subgoaling reduces the complexity of probabilistic inference and problem solving. *J. R. Soc. Interface* 12:20141335. doi: 10.1098/rsif.2014.1335

Mayer, R. C., Davis, J. H., and Schoorman, F. D. (1995). An integrative model of organizational trust. *Acad. Manag. Rev.* 20, 709–734. doi: 10.2307/258792

Méndez, J. C., Pérez, O., Prado, L., and Merchant, H. (2014). Linking perception, cognition, and action: psychophysical observations and neural network modelling. *PLoS One* 9:e102553. doi: 10.1371/journal.pone.0102553

Morry, M. M. (2005). Relationship satisfaction as a predictor of similarity ratings: a test of the attraction-similarity hypothesis. *J. Soc. Personal Relationships* 22, 561–584. doi: 10.1177/0265407505054524

Moutoussis, M., Fearon, P., El-Deredy, W., Dolan, R. J., and Friston, K. J. (2014). Bayesian inferences about the self (and others): a review. *Conscious Cogn.* 25, 67–76. doi: 10.1016/j.concog.2014.01.009

Muir, B. M. (1994). Trust in automation: Part I. theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics* 37, 1905–1922. doi: 10.1080/00140139408964957

Nass, C., Moon, Y., Fogg, B. J., Reeves, B., and Dryer, D. C. (1995). Can computer personalities be human personalities? *Int. J. Human-Computer Stud.* 43, 223–239. doi: 10.1006/ijhc.1995.1042

Nave, K., Deane, G., Miller, M., and Clark, A. (2020). Wilding the predictive brain. *Wiley Interdisciplinary Rev. Cogn. Sci.* 11:e1542.

Norman, D. A., Ortony, A., and Russell, D. M. (2003). Affect and machine design: lessons for the development of autonomous machines. *IBM Systems J.* 42, 38–44. doi: 10.1147/sj.421.0038

Parr, T., and Friston, K. J. (2017). Uncertainty, epistemics and active inference. *J. R. Soc. Interface* 14:20170376. doi: 10.1098/rsif.2017.0376

Parr, T., and Friston, K. J. (2019). Generalised free energy and active inference. *Biol. Cybern.* 113, 495–513. doi: 10.1007/s00422-019-00805-w

Parr, T., Rees, G., and Friston, K. J. (2018). Computational neuropsychology and Bayesian inference. *Front. Hum. Neurosci.* 12:61. doi: 10.3389/fnhum.2018.00061

Perlovsky, L., and Schoeller, F. (2019). Unconscious emotions of human learning. *Phys. Life Rev.* 31, 257–262. doi: 10.1016/j.plrev.2019.10.007

Pessoa, L. (2010). Emotion and cognition and the amygdala: from "what is it?" to "what's to be done?". *Neuropsychologia* 48, 3416–3429. doi: 10.1016/j.neuropsychologia.2010.06.038

Peters, A., McEwen, B. S., and Friston, K. (2017). Uncertainty and stress: why it causes diseases and how it is mastered by the brain. *Prog. Neurobiol.* 156, 164–188. doi: 10.1016/j.pneurobio.2017.05.004

Pio-Lopez, L., Nizard, A., Friston, K., and Pezzulo, G. (2016). Active inference and robot control: a case study. *J. R. Soc. Interface* 13:20160616. doi: 10.1098/rsif.2016.0616

Pisula, W., and Siegel, J. (2005). Exploratory behavior as a function of environmental novelty and complexity in male and female rats. *Psychol. Rep.* 97, 631–638. doi: 10.2466/pr0.97.2.631-638

Ramstead, M. J., Wiese, W., Miller, M., and Friston, K. J. (2020). *Deep Neurophenomenology: An Active Inference Account of Some Features of Conscious Experience and of their Disturbance in Major Depressive disorder*. Available online at: http://philsci-archive.pitt.edu/18377/ (accessed 30 April, 2021)

Rao, R. P., and Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.* 2, 79–87. doi: 10.1038/4580

Rempel, J. K., Holmes, J. G., and Zanna, M. P. (1985). Trust in close relationships. *J. Pers. Soc. Psychol.* 49:95.

Salge, C., and Polani, D. (2017). Empowerment as replacement for the three laws of robotics. *Front. Robot. AI* 4:25. doi: 10.3389/frobt.2017.00025

Salomon, R. (2017). The assembly of the self from sensory and motor foundations. *Soc. Cogn.* 35, 87–106. doi: 10.1521/soco.2017.35.2.87

Salomon, R., Fernandez, N. B., van Elk, M., Vachicouras, N., Sabatier, F., Tychinskaya, A., et al. (2016). Changing motor perception by sensorimotor conflicts and body ownership. *Sci. Rep.* 6:25847. doi: 10.1038/srep25847

Salomon, R., Progin, P., Griffa, A., Rognini, G., Do, K. Q., Conus, P., et al. (2020). Sensorimotor induction of auditory misattribution in early psychosis. *Schizophrenia Bull.* 46, 947–954. doi: 10.1093/schbul/sbz136

Schoeller, F. (2015). Knowledge, curiosity, and aesthetic chills. *Front. Psychol.* 6:1546. doi: 10.3389/fpsyg.2015.01546

Schoeller, F. (2017). The satiation of natural curiosity. *Int. J. Signs Semiotic Systems* 5, 200516–232707.

Schoeller, F., and Perlovsky, L. (2016). Aesthetic chills: knowledge-acquisition, meaning-making and aesthetic emotions. *Front. Psychol.* 7:1093. doi: 10.3389/fpsyg.2016.01093

Schoeller, F., Perlovsky, L., and Arseniev, D. (2018). Physics of mind: experimental confirmations of theoretical predictions. *Phys. Life Rev.* 25, 45–68. doi: 10.1016/j.plrev.2017.11.021

Schoeller, F., Eskinazi, M., and Garreau, D. (2017). Dynamics of the knowledge instinct: effects of incoherence on the cognitive system. *Cogn. Systems Res.* 47, 85–91. doi: 10.1016/j.cogsys.2017.07.005

Seth, A. (2014). "The cybernetic brain: from interoceptive inference to sensorimotor contingencies," in *MINDS project*, eds T. Metzinger and J. M. Windt (Glastonbury, CT: MINDS).

Seth, A. K. (2013). Interoceptive inference, emotion, and the embodied self. *Trends Cogn. Sci.* 17, 565–573. doi: 10.1016/j.tics.2013.09.007

Seth, A. K. (2015). "Inference to the best prediction," in *Open MIND*, ed. J. M. Windt (Glastonbury, CT: MIND Group).

Seth, A. K., and Friston, K. J. (2016). Active interoceptive inference and the emotional brain. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 371:20160007. doi: 10.1098/rstb.2016.0007

Sheridan, T. B. (1988). "Trustworthiness of command and control systems," in *Proceedings of IFAC Man-Machine Systems*, (Oulu), 427–431.

Sheridan, T. B. (2019a). Extending three existing models to analysis of trust in automation: signal detection, statistical parameter estimation, and model-based control. *Hum. Factors J. Hum. Factors Ergonom. Soc.* 61, 1162–1170.

Sheridan, T. B. (2019b). Individual differences in attributes of trust in automation: measurement and application to system design. *Front. Psychol.* 10:1117. doi: 10.3389/fpsyg.2019.01117

Sheridan, T. B., and Parasuraman, R. (2005). Human-Automation interaction. *Rev. Hum. Factors Ergonom.* 1, 89–129.

Smith, R., Parr, T., and Friston, K. J. (2019). Simulating emotions: an active inference model of emotional state inference and emotion concept learning. *Front. Psychol.* 10:2844. doi: 10.3389/fpsyg.2019.02844

Stephan, K. E., Manjaly, Z. M., Mathys, C. D., Weber, L. A. E., Paliwal, S., et al. (2016). Allostatic self-efficacy: a metacognitive theory of dyshomeostasis-induced fatigue and depression. *Front. Hum. Neurosci.* 10:550. doi: 10.3389/fnhum.2016.00550

Synofzik, M., Vosgerau, G., and Newen, A. (2008). Beyond the comparator model: a multifactorial two-step account of agency. *Conscious Cogn.* 17, 219–239.

Van de Cruys, S. (2017). *Affective Value in the Predictive Mind*. Frankfurt: MIND Group.

Veissière, S. P. L., Constant, A., Ramstead, M. J. D., Friston, K. J., and Kirmayer, L. J. (2019). Thinking through other minds: a variational approach to cognition and culture. *Behav. Brain Sci.* 43:e90. doi: 10.1017/S0140525X19001213

Vuorre, M., and Metcalfe, J. (2016). The relation between the sense of agency and the experience of flow. *Conscious Cogn.* 43, 133–142.

Wilkinson, S., Deane, G., Nave, K., and Clark, A. (2019). "Getting warmer: predictive processing and the nature of emotion," in *The value of Emotions for Knowledge* ed. L. Candiotto (Cham: Palgrave Macmillan), 101–119.

Wilson, M., and Wilson, T. P. (2005). An oscillator model of the timing of turn-taking. *Psychon. Bull. Rev.* 12, 957–968. doi: 10.3758/bf03206432

Wolpert, D. M., Ghahramani, Z., and Jordan, M. I. (1995). An internal model for sensorimotor integration. *Science* 269:1880.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer AO declared a past collaboration with one of the authors FS to the handling Editor.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.