PHILOSOPHICAL TRANSACTIONS B

royalsocietypublishing.org/journal/rstb

Opinion piece

Cite this article: Kahl S, Kopp S. 2023 Intertwining the social and the cognitive loops: socially enactive cognition for human-compatible interactive systems. *Phil. Trans. R. Soc. B* **378**: 20210474. https://doi.org/10.1098/rstb.2021.0474

Received: 14 July 2022
Accepted: 9 January 2023

One contribution of 15 to a discussion meeting issue 'Face2face: advancing the science of social interaction'.

Subject Areas:
cognition

Keywords:
face-to-face interaction, interactive intelligent systems, socially enactive cognition, predictive processing, online mentalizing

Author for correspondence:
Stefan Kopp
e-mail: skopp@techfak.uni-bielefeld.de

THE ROYAL SOCIETY PUBLISHING

Intertwining the social and the cognitive loops: socially enactive cognition for human-compatible interactive systems

Sebastian Kahl and Stefan Kopp

Social Cognitive Systems Group, Faculty of Technology and CITEC, Bielefeld University, 33619 Bielefeld, Germany

SKa, 0000-0002-8468-2808; SKo, 0000-0002-4047-9277

It is increasingly important for technical systems to be able to interact flexibly, robustly and fluently with humans in real-world scenarios. However, while current AI systems excel at narrow task competencies, they lack crucial interaction abilities for the adaptive and co-constructed social interactions that humans engage in. We argue that a possible avenue to tackle the corresponding computational modelling challenges is to embrace interactive theories of social understanding in humans. We propose the notion of socially enactive cognitive systems that do not rely solely on abstract and (quasi-)complete internal models for separate social perception, reasoning and action. By contrast, socially enactive cognitive agents are supposed to enable a close interlinking of the enactive socio-cognitive processing loops within each agent, and the social-communicative loop between them. We discuss theoretical foundations of this view, identify principles and requirements for according computational approaches, and highlight three examples of our own research that showcase the interaction abilities achievable in this way.

This article is part of a discussion meeting issue 'Face2face: advancing the science of social interaction'.

1. Introduction

Modern AI (artificial intelligence) systems are capable of achieving astonishing performances in a variety of tasks, e.g. classifying objects in video images, solving complex control problems, or turning language descriptions into artistic visualizations. But to leverage these abilities in many socially important domains, AI systems also have to master real-world encounters with humans. For example, autonomous vehicles need to navigate traffic situations involving human pedestrians or car-drivers, collaborative robots are supposed to work alongside human workers on assembly lines, or personal assistants should give recommendations for travel or healthy living and explain those to individual users. Thus, it is crucial for intelligent systems to be able to engage in extended, meaningful interactions to support, assist, cooperate with, or learn from humans.

A large body of work has been directed to enabling interaction of humans with machines, robots or now AI technology, and dedicated sub-fields specialized in finding technological and design solutions to this. Early work in those fields has focused on making technical systems 'ready for use', with a human user being in full control and in charge of driving the task towards a solution. With the advent of intelligent autonomous systems with abilities complementing or exceeding those of humans, this focus has shifted towards making systems 'ready for cooperative interaction' with humans. Yet, equipping AI systems with the required *social interaction intelligence*, in addition to their task intelligence, remains a key challenge [1].

The literature on robotics and artificial social agents increasingly argues that systems need to be more 'aware' of their human user, the specific situation, or the larger social context. Correspondingly, there is a growing number of proposals to make AI systems context-aware, socially aware or even human-aware [2]. The term social awareness has been used to denote a machine's interpretation

of the social context that is inferred from detected signals and taken to provide the necessary information to appropriately interact with peers [3]. Human-awareness is used to refer to the development of systems that can understand the mental states of humans, provide proactive support, or give cogent explanations of their actions [2]. Technically speaking, this work involves developing computational methods to recognize contextually significant behavioural or situational features, infer latent world states, or determine action decisions that reconcile task-specific needs with effects on a human user. For example, approaches to the so-called 'human-aware planning' [4] try to equip robots with an approximate model of what a human user wants or knows about the environment, and then to respond to this either by generating behaviour that is readily understandable (explicable) to the user, or by deriving explanations that would alter the human's model and hence make the robot's action more understandable.

Despite this progress in promising and important directions, a core ingredient for artificial systems to be truly 'interaction-ready' in a human-compatible form is still missing. Existing approaches aim to identify and solve the perception, reasoning, and decision-making problems in interactions with other agents as if they were similar to those in non-social tasks. But controlling a robot in a factory environment is different from communicating with an elderly user to figure out what he/she needs and provide support that she understands and accepts. Yet, although the problems are fundamentally different, current approaches are similar and focus on training specialized models that, e.g. recognize a user's emotional state or degree of understanding from social cues, or describe action policies that maximize predetermined task-level or social cost functions. The shortcomings of this approach arise from the

— separation of social perception, reasoning, decision-making and action in order to come up with modular, specialized solutions for these (in fact, inseparable) sub-problems. As a result, current systems excel, for instance, in social signal processing or natural language conversation but get complex and clumsy when plugging these heavy-weight modules together.
— neglect of the subjective mental perspectives of the interacting agents that come to be mutually established and extended through social interaction. While previous work has tried to formalize the relevant internal states of human users and interactive agents, their mental perspectives towards each other's mental states along with a suitable notion of *sharedness* of goals, beliefs or attitudes still have not been accounted for sufficiently.
— neglect of predictive and embodied cognitive processes in social interaction. As a result, current systems rely on complex internal models to process rich information, but are hardly able to respond fluently or align quickly to verbal and nonverbal behaviour of human users.
— neglect of (i) the dynamics of socio-cognitive processes within the interactants' minds, (ii) the dynamics of the social interaction and the couplings between them, and (iii) the interplay between these two kinds of closed-loop dynamics. As a result, systems are not able to co-adapt continuously with individual users over the course of a single interaction taking place in a specific situational context.
— detachment of the construction/learning of social signals or communicative behaviour from their use in online interaction, during which they would need to be continuously modulated or adapted to meet the specific needs of the interactants and their situational context.

We argue that one approach to overcome these limitations is by building artificial intelligent agents that incorporate principles of 'socially enactive cognition'. By this we mean cognitive systems that are not solely based on 'cognitivist' computations over complex representational models of social situations inside an agent. By contrast, we emphasize the view that artificial interactive systems *additionally* need to be equipped with perceptual and socio-cognitive abilities that are grounded in dynamic meaningful relationships arising from an adaptive two-way exchange between the systems and the social environment they inhabit and actively shape [5]. The goal is to build artificial systems that can engage in interactions in which a *social loop* couples interaction partners via dynamical reciprocal processes of communication and coordination, while being bi-directionally interlinked with the intra-agent *cognitive loops* such that this two-way exchange is meaningfully effective within and between the agents. We conjecture that this can yield an interaction quality that makes artificial intelligent systems significantly more human-compatible.

In the following, we formulate possible avenues for tackling the corresponding modelling challenge by embracing interactive theories of social understanding. We start by discussing theoretical foundations, identify important tenets for respective computational approaches, and highlight examples of our own research that showcase the interaction abilities achievable in this way. Finally, we will discuss missing steps towards intelligent systems that can engage in human-compatible interaction and cooperation.

2. Socially enactive minds for interactive intelligent systems

What are socially enactive cognitive systems and how can we go about building interactive systems, artificial assistants or robots based on these principles? We first clarify what we mean by socially enactive minds, by discussing relevant theoretical and empirical foundations. Based on this, we will propose corresponding computational principles and approaches for building artificial systems that exhibit similar qualities in social interactions with human users.

(a) Theoretical foundations: enactive cognition in social minds

A large body of research has been directed to studying the phenomenology of human social interactions and the human abilities for social cognition. But only in the last decade have researchers started to realize the importance of studying both together [6,7]. Up to then investigation of social cognition was mostly focused on individual mechanisms, mainly considering the third-person perceptual and inferential processes of determining subjective mental states of other agents (their beliefs, intentions, goals, attitudes, etc.). This so-called mentalizing or Theory of Mind (ToM) [8] requires an observer to have a good model of the acting agent and has traditionally been described as a form of theory building or simulation. Theory-theory

(TT) views ToM as a separate system that enables reasoning about (folk) psychology—a framework of concepts that allow explanations to be derived from a set of laws and rules that '[···] connect the explanatory conditions with the behaviour explained' [9, p. 68]. Simulation theory (ST), in contrast, assumes that we do not have a separate black-box system to answer psychological queries. Rather, we reason about ourselves to simulate the minds of others. Most often this process is described as a form of simulation where the own social perception apparatus is taken *offline* to run simulations with pretend mental states, to reach conclusions about the other agent's behaviour [10].

These classical views of social understanding have been criticized substantially. For one, they focus on social perception or action tasks but these have been studied mainly in isolation. Correspondingly, these paradigms assume that social interaction is reducible to the understanding of mental states that people entertain when they interact [11] and that those mental states are 'hidden, pre-existing, unaffected by the interaction, owned by each participant and opaque to the other' [12, p. 4]. Further, when adopted to explain how humans understand other minds, these paradigms subscribe to the largely refuted 'sandwich model', according to which socio-cognitive processes lie in-between perception (as input from the world to the mind) and action (as output of the mind) [13].

Notions such as direct social perception, enactive cognition and participatory sense-making [14] have been put forward as alternatives. Direct social perception refers to the view that socially relevant mental states are often directly perceivable (e.g. motor intentions from motion) and thus need not be inferred. This relies heavily on a perception that is *smart* in a way that allows the individual to extract what usually are described as *hidden* mental states, and that is 'innately (or very early) tuned to socially relevant aspects of the world' [15, p. 540].

On a more general note, embodied cognitive views have rejected the sole importance of internal mental representations for many tasks. In particular, enactive cognition assumes that cognition in the individual is constituted by sensorimotor activity [16] and that perception is *active sense-making* that arises from embodied engagement with the environment and prepares for action. That is, just like an organism is driven by the need to maintain biological balance, cognition is shaped by a form of autopoietic interaction with the information environment [17,18]. The underlying notion of self-construction of an individual's reality was applied to social reality early on, most prominently in Bruner's theory of instruction [19] and Vygotsky's social constructivism, according to which people learn internal social reasoning from external social interaction [20].

Extending the enactive view to social cognition leads to understanding social interaction—contingent perception and action in a social situation—as being essential for the cognitive capacity of understanding others. This view has led to the *interactive brain hypothesis* (IBH) [12,21], stating that social understanding capabilities are a result of the skilful use of social interaction capabilities acquired during an individual's upbringing. The IBH was put forward from an enactivist perspective focusing on autonomy, (participatory) sense-making and embodiment [22] with a number of assumptions: first, our skills constituting social understanding are developed through our experience of social interaction, developmentally predisposed by a readiness to interact [12] and to attend others [23]. Secondly, social interaction is a contextual factor, where meaning that becomes only apparent in interaction can influence the meaning of individual presumptions. Thirdly, processes in the *individual* brain during social interaction are *not* fully constitutive of social cognition [21], rather the acts and meanings cognized during social interaction are irreducible and emergent properties of the interaction: 'To cognize socially, in the enactive understanding of the term, is to skilfully engage in the multiple demands and possibilities of the social world, many of which are directly or indirectly emergent from social interactions' [21, p. 6]. The autonomy aspect of the IBH relates to the autopoietic view and makes it compatible with a predictive processing perspective according to which an autonomous agent brings forth its own action–effect couplings with the environment in a sense-making process that strives to minimize prediction errors [24].

Despite the potential of enactivist positions (such as IBH) for improving artificial social systems (see below), it is important to note that radical views of enactive cognition, which ultimately reject notions of mental states and representations in cognition altogether, can hardly account for one basis of communication and social interaction, namely, the construal and sharing of information about non-present or even imaginary entities. We underscore the role of mental representations for a cognitive system to be able to establish social understanding. In particular, the ability to differentiate between subjective mental perspectives, such as *my* beliefs versus *your* beliefs or our common *we-beliefs*, is essential for effectively tracking the (non-)success of social understanding. This is a notion of implicit mentalizing where the *we-mode* [25] or *we-beliefs* [26,27] are a form of sharedness that helps to bootstrap the commonly assumed recursive nature of mental state attribution.

In sum, we ground the idea of artificial, socially enactive cognitive systems upon theories of autopoietic enactive cognition in the domain of social understanding and interaction [17]. As we will specify below, social understanding in such systems results from a bottom-up process of interactive sense-making, grounded in dynamic interaction with the social world, and a top-down process of ongoing skilful interaction enrichment through the individual's experience with social interactions as well as the individual interaction partner.

(b) Towards artificial socially enactive minds

The hallmark and linchpin of socially enactive artificial systems is their ability to engage in a social interaction in which the dynamics of enactive socio-cognitive processes (within interactants) and the communicative–coordinative processes (between them) unfold in parallel and in close relation to each other. Crucially, this relation is seen as bi-directionally constitutive. On the one hand, social understanding and cognition should be directly grounded in, and to some extent arise from, the running social interaction. On the other, social interaction, and its intricate complexity of various communicative cues and signals being produced and perceived simultaneously, are driven by the underlying socio-cognitive processes of the interactants. To answer how this quality could be attained in artificial systems, one must address this interplay from both sides: by asking what dynamic interaction phenomena are characteristic of the social loops we are aiming for, and by asking how socially

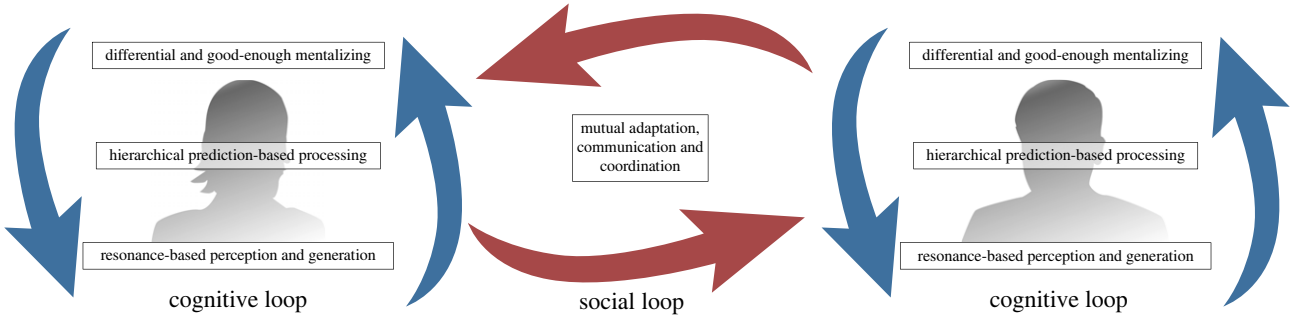


Figure 1. Socially enactive cognitive systems engage in two coupled dynamic processes: the social loop (red arrows) consists of mutual adaptation, communication and coordination processes between the agents. It arises from, and is constitutive of cognitive loops (blue arrows) that unfold within each agent. Each cognitive loop links resonance-based perception and generation of social signals with differential mentalizing processes through hierarchical prediction-based processing (for explaining and predicting one another's behaviour). This allows interlinking of online social cognition and social interaction, e.g. by creating and testing mental state attributions through reciprocal coordination and alignment processes. (Online version in colour.)

enactive minds operate in this and how it can be modelled computationally.

As regards the social loop dynamics, elsewhere we have discussed the various forms of inter-agent coordination and mutual adaptations that constitute a socially resonant interaction [28]. The main degrees of freedom of these processes result from the ability of socially enactive systems to perceive and produce communicative signals, such as speech, prosody, gaze, facial expression, gesture, body posture, etc. (cf. [29]). This involves behaviours transferred via different sensory modalities (e.g. verbal, vocal, non-vocal), serving different functions (e.g. to convey content, emphasize, demonstrate emotions, provide feedback, negotiate turns) with different degrees of intentionality and awareness (e.g. unaware indicators, intended displays, intended-to-be-recognized-as-displayed signals [30]) and often as part of a larger multimodal delivery. Through the adaptive use of these *polysemiotic communicative systems*, human interlocutors have been found to mutually coordinate and socially resonate in various ways (cf. [28]): from a strategic convergence on verbal-acoustic styles (as in Speech Accommodation Theory [31]) or the joint activity of grounding linguistic constructions [32], to a priming-based alignment of prosodic or lexico-grammatical choices [33], to an automatic mimicry [34] or temporal entrainment of nonverbal behaviour [35].

It is noteworthy that these mutual adaptation and coordination processes, though sometimes being implemented individually and explicitly (cf. [36]), are not yet found as emerging systematically in embodied human–agent interaction. We argue that this is due to the lack of the ability of artificial systems to operate in socially enactive cognitive ways. Figure 1 illustrates this view of interlinked cognitive and social loops. Note that we do not require (nor assume) perceptual, cognitive, affective or motor processes to be identical in humans and artificial systems. What we do argue, however, is that these intra-agent processes incrementally give rise to and arise from multimodal signals that are communicated for the purpose of mutual adaptation and coordination, and that artificial systems 'ready for cooperative interaction' must support *both* kinds of dynamic processes for their human users. To that end, they must be built with corresponding *guiding principles* in mind. As will be explained next based on work of others as well as ourselves, we conceive of those as perceptual or motor resonances that result from incremental, hierarchical prediction-based processes on (possibly false, but repairable) representations of conditional mental state attributions.

(i) Resonance-based behaviour perception and generation

One guiding principle for building artificial agents that support a more enactive approach to social understanding looks into the mechanistic underpinnings of action understanding in social encounters. From studies of the neural basis of action understanding, so-called *motor resonances* are known to occur in the brain of action observers [37]. One source of explanations for such resonances is mirror neuron activity, which is a strong indicator for the involvement of the motor system in deriving predictions and evaluating hypotheses about incoming observations. In humans, mirror neurons have been found in different parts of the brain [38,39] and a debate has arisen around whether their origin is either evolutionary or a form of associative learning [40]. For theories of behaviour understanding, they have spawned discussions about the so-called perception–action links, common coding [41] and event coding [42]. In addition, from a dynamic systems perspective, correlated neural activation found in parts of the brain of interacting humans fits well with an embodied cognition view where our nervous system is coupled with the environment through our bodies and bodily senses. In social environments, one's social acts, perceived by all participants of an interaction, lead to sympathetic resonances in their respective nervous systems and hence attune the way they interact. Recent evidence even suggests that interacting individuals show interpersonal coupling or even alignment of resonating neural dynamics underlying cognitive and affective processes [43,44]. While the actual mechanisms and effects of this kind of coupling are still not fully understood, it adds evidence for the view that automatic resonances of perceptual and motor processes may be a basis for inter-agent adaptation and coordination. It is also in line with classical accounts such as Adaptive Resonance Theory [45], which introduced predictive coding as underlying computational mechanism and led to a Bayesian perspective on the mirror neuron system as a source for social understanding in work on predictive processing [46,47].

(ii) Hierarchical prediction-based processing

A second guiding principle that computational models should try to follow is to overcome the strict separation of

input processing and behaviour generation, as well as the dichotomy of lower-level and higher-level processing. De Bruin & Strijbos [48] argue that principles of predictive processing and, in particular, the so-called Free Energy Principle (FEP) provide a perspective toward social cognition that overcomes the classical sandwich-model-based views on social understanding and social interaction. The FEP was developed as a theory for perception under the framework of predictive coding, wherein probabilistic distributions over latent variables in generative models are inferred for the observed sensation by minimizing the prediction error (the free energy) [24,47,49].

Later, a corollary of the FEP, active inference, was developed as a framework for action generation wherein action policies as well as the probabilistic distribution of future latent variables in generative models are inferred by minimizing the so-called expected free energy (uncertainty with respect to goal-directed action outcomes). The major advantage of this approach for goal-directed action, compared with conventional forward models, is that the model can cope with uncertainty in a stochastic environment by incorporating a Bayesian framework. Active inference describes a process of decision-making and action production in which the environment is affected to reduce uncertainty about an agent's beliefs about the world [50]. As in the Bayesian brain perspective, where predictions about sensory stimuli are continuously formed and evaluated in a generative process, this generative process is inverted to predict next actions and thus attenuates prediction errors [51]. This is an affordance competition process, i.e. action selection is based on possible goals achievable through that action [52].

The FEP gets rid of the *perception–cognition* distinction by seeing perception as unconscious inference (going back to Helmholtzian perspectives on perception). Also it accepts perception to be driven by expectations, i.e. priors that are constantly formed in the brain and contrasted against perceived sensory evidence. Only the interesting (information-carrying) aspects of the input, the prediction errors, are passed up the hierarchy for further (more abstract) processing. As a result, this account mitigates the *perception–action* distinction as we no longer need constant deep inferential processing in a processing hierarchy if the input is already expected. Ideas of direct social perception, which claim that we can directly perceive some of our interaction partner's mental states, have also been based on this view [15].

(iii) Differential, good-enough mentalizing

A third key principle for interactive systems is the ability to differentiate between subjective mental states of the interaction partners (i.e. their subjective beliefs, intentions, goals, attitudes, etc.). This amounts to forming an *online* theory of mind for other agents, which is necessary to sufficiently characterize the state of an interaction between cognitive agents (e.g. cooperative or competitive) to propel communication forward in order to, e.g., extend or adjust another agent's knowledge, or identify and repair communication problems. We particularly emphasize the importance of the latter, i.e. a goal-directed and flexible management of uncertainties of (mis-)understandings. For this to be possible in human–machine interaction settings, artificial systems need deep (as opposed to shallow) models of the cognitive processes that play out during social interaction. Yet, today's AI systems are only able to account for these aspects offline and in narrow toy scenarios.

Classical approaches to dialogue modelling have looked into formal models of the so-called conversational records or information states [53,54], which describe the information that is currently private or shared (grounded) between interlocutors. However, a socio-cognitive approach to distinguish mental perspectives is rarely adopted. For one thing, this is due to the fact that modelling nested beliefs and trying to infer them as hidden states gets complex and quickly even intractable. Thus, we argue for artificial socio-cognitive systems to follow a good-enough or 'satisficing' approach to mentalizing, by employing minimal ToM models when possible and more complex ones when needed [55]. For example, a minimal model would differentiate between three perspectives: my own beliefs (first-person), the interlocutor's beliefs (second-person) and our shared beliefs (first-person plural). The latter corresponds to mental states that *every* interactant believes to hold for *all* interactants, including oneself. This model was shown to be able to account for basic initiation as well as repair phenomena found in social interaction [26]. In the model implementation examples presented below, the first example describes a model where prediction errors during inference of social information are accounted for as false beliefs, which leads to repairs and other reciprocal belief coordination behaviour through communication [27], a notion similar to what has even been called the running repair hypothesis [56]. More complex problems of miscommunication, however, like figuring out that the other has expected you to say something else or disagrees, demand more complex social reasoning and mutual coordination processes (as shown in the second example).

Friston & Frith [57] describe a solution to the ToM. This solution replaces the problem of inferring another's mental state with inferring what state of mind one would be in to produce the same sensory consequences (similar to ideomotor theory). The basic idea is that internal or generative models used to infer one's own behaviour can be deployed to infer the beliefs (e.g. intentions) of another, assuming both parties have sufficiently similar generative models. Indeed, previous literature discusses a *readiness to interact*, a *we-mode*, *we-belief* or a *prior belief* about the similarity of conspecifics [12,25,26,58,59], which all can be interpreted and modelled as an 'abductive bias' during mentalizing.

(iv) Incremental processing loops

A final and increasingly adopted principle is incremental processing, both in the social loop and in the cognitive loop. Previous work has shown that incrementality and the responsiveness it enables improve the fluidity and smoothness of conversational interaction (cf. [60]). For example, speakers are producing their verbal utterance in a step-wise fashion while being able to re-plan remaining part(s) of their communicative plan, suspend it, inject a sub-dialogue and continue at a later point in ways adapted to the addressees' needs. They can do so in such an effortless, smoothly coordinated and seemingly natural way that it is not even apparent that difficulties were paid attention to or that plans were changed mid-way. It is important to note, however, that incremental processing goes well beyond producing behaviour step-by-step. In particular, it requires the cognitive loop to be able to run simultaneously at multiple levels of a processing hierarchy, often operating on

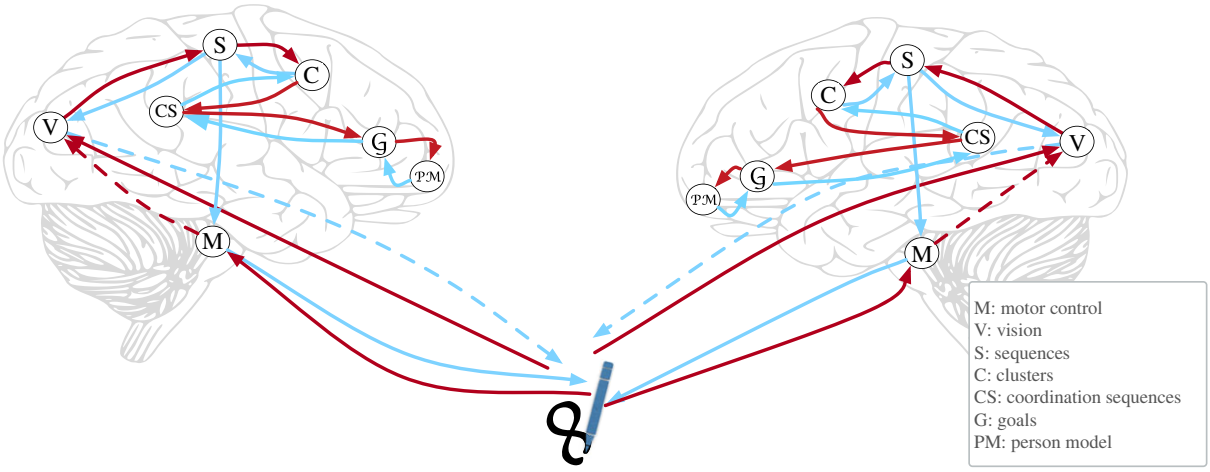


Figure 2. Two 'agents' interact with each another by reciprocally perceiving and performing the handwriting of digits. Each agent is based on a hierarchy of nested generative processes, spanning main sensorimotor levels (M, V, S, C) as well as levels associated with mentalizing (CS, G, PM). Across these levels, the generative processes predict the activity of the next-lower level, while prediction errors determined from visual input (V) and proprioceptive feedback traverse the hierarchy back upwards. By coupling these models through their interaction, agents reciprocate by writing what they believe they have understood and that way coordinate their beliefs dynamically and incrementally until they reach a mutual understanding. (Online version in colour.)

premature hypotheses and being able to quickly adapt when they are updated based on incoming sensory input. Such an incremental processing naturally follows from the hierarchical predictive-processing principle, when *partial* (bottom-up) interpretation hypotheses are combined with the evaluation of (top-down) predictions about, e.g., an interlocutor's next contribution. Taking this further to socially enactive processing means that a system is not only able to process and produce socio-communicative behaviour incrementally, but also to do this at different multi-layered coordination loops, which are maintained with one another to establish mutuality of, e.g., contact, understanding or goals.

3. Example implementations and results

In the following, we will present three examples of actually implemented computational systems that embody (parts of) a socially enactive approach to social understanding and interaction. For each of them, we outline the scenario, the implemented modelling approach, which of the above principles they are based on, and results obtained in this way.

(a) Online predictive processing of communicative behaviour

We have explored how predictive processing can foster online social interaction. Two artificial social agents were independently equipped with a computational model that dynamically forms predictive beliefs during the online perception and production of communicative signals (here, the writing of digits). Based on these beliefs, the agents were enabled to reciprocally react to each other in order to reach a joint (mutually aware) understanding of a digit [27,61]. The model consists of a hierarchy of nested generative sensorimotor perception and action processes and is based on the free-energy principle with active inference [46,47,51]. Active inference sees action as a form of inference over possible ways to make the environment meet the agent's predictions, here about the inferred mental states of the social interaction partner.

See figure 2 for a sketch of two models interacting with one another through their perceiving and acting of the handwriting of digits. Each model consists of a nested hierarchy of specialized generative processes that each predicts the activity of its next-lower level, where predictions of the motor control model (M) lead to overt action. The success of predictions as well as prediction errors evaluated on visual input (V) and proprioceptive feedback traverse the hierarchy upwards towards levels that are associated with mentalizing (CS, G). This enables a process for belief coordination similar to the *we-mode* proposed by Frith [62]: an implicit form of mentalizing that bootstraps the attribution of mental states, which seems automatic, where contextual information and prior information can influence behavioural understanding top-down.

We argue that nested high-level representations that enable belief coordination do not only act as attentional mechanisms to minimize uncertainty during perception and action. They also assume a form of meta-cognitive regulatory control based on the uncertainty detected in the social interaction overall. That is, they can orchestrate the perception and production of social signals and are vital in resolving uncertainty in belief coordination [63]. In other words, prediction errors during inference of social information may be inferred as false beliefs and lead to repairs and other reciprocal communication behaviour that is integral to communication [56]. Technically, the hierarchy performs a general belief-update at each level, based on a linear Bayesian process that involves a dynamic information gain (or 'precision-weighting') related to the system's uncertainty. This precision-weighting can be biased by more strategic, meta-cognitive regulation mechanisms to allow weighted information gain and to affect the agent's perception, from focusing on details to ignoring detailed prediction errors.

In simulations with three such agents in a nonverbal communication game, we have shown that agents perform reciprocal belief coordination that involves regulation of the whole-nested hierarchy with sensorimotor and mentalizing parts. The results stress the importance of balancing prior beliefs against new information by means of a precision-

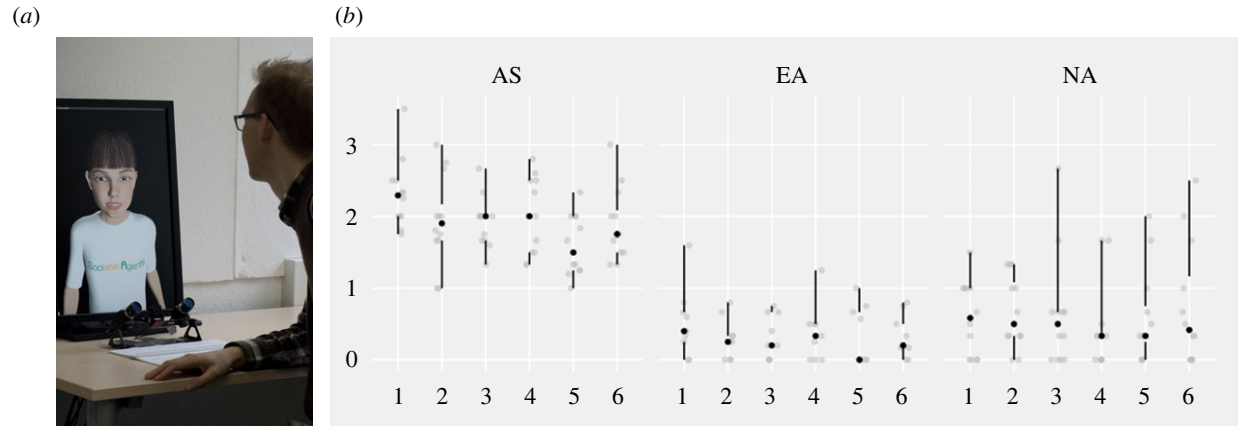


Figure 3. (*a*) An 'attentive speaker' agent communicates with a human user while interpreting the interlocutor's communicative feedback and adapting to it online. (*b*) Results demonstrate that humans engage in this and produce significantly more feedback with the attentive speaker (AS) than with agents that do not attend (NA) or explicitly ask for confirmation all the time (EA) [67]. Datapoints are light grey, black dots are medians, black lines are whiskers representing 1.5 × interquartile range, and mid gaps are quartiles. (Online version in colour.)

weighting bias that regulates not only perception and action, but also more generally the gain on social information.[1]

This work embodies several of the aforementioned principles for modelling socially enactive artificial systems. Most notably, it implements *resonance-based behaviour perception and generation* and an *hierarchical prediction-based processing* of social representations from behaviour to mental states, in the form of a hierarchical generative model. This allows the agent to act upon prediction errors about high-level beliefs about the interaction partner's beliefs, entailing either reciprocating inference of beliefs or attempts to repair false beliefs as a form of active social inference. In this sense, the model grounds its decision-making on the principle of *differential and good-enough mentalizing* as it infers beliefs that are good enough for fluent interaction, in contrast to slow and resource-intensive optimal decision-making. Finally, regarding the principle of *incremental processing*, the implemented model processes and produces social behaviour incrementally based on partial hypotheses underlying current predictions. However, while strategies for handling false beliefs are chosen as they occur, the model is not able to self-repair and coordinate the complete hierarchy of nested loops in the process.

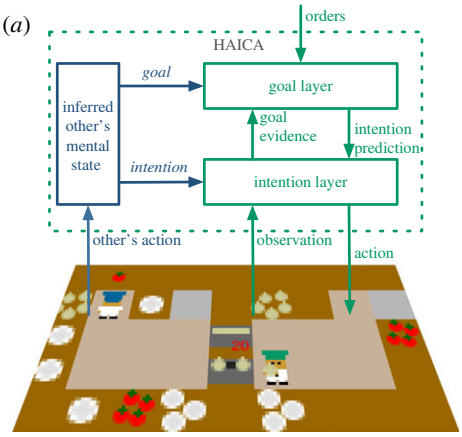
(b) Adaptive coordination of dialogue through feedback and mentalizing

Spoken-language or multimodal dialogue is a 'collaborative effort' of the interlocutors who cooperatively try to establish mutual understanding [32,64]. One prevalent mechanism that humans use in such situations is listener feedback [65], short unobtrusive signals that listeners produce in the 'back-channel', while processing verbal input produced by speakers that in turn perceive this feedback and adapt their own other-directed behaviour accordingly. This is an example *par excellence* for socially enactive processing in artificial systems. Feedback signals are employed to dynamically coordinate an incremental social interaction, shaped by the dynamic socio-cognitive processing of the listener, driving the corresponding processes in the speaker, which in turn shapes his/her incremental utterance, which again affects the socio-cognitive processes in the listener. The social and cognitive loops in this example rely on different (verbal and nonverbal) modalities, along with corresponding sensorimotor processes grounded in the embodiment of the conversation partners.

Buschmeier & Kopp [66] have presented a model that enables feedback-based coordination between an embodied conversational agent and a human interaction partner (figure 3*a*). The implemented 'attentive speaker agents' are able (1) to interpret communicative listener feedback of their human interlocutors, taking the dialogue context into account, and (2) to adapt their ongoing or subsequent natural language utterances according to their interlocutors' needs—as inferred in (1). For this, the agent builds an internal model of the listening-related mental state of the human user, and updates it continuously and incrementally based on the perceived feedback signals. That is, the model is strongly based on the principle of *incremental processing*. In addition, it follows the principle of *good-enough mentalizing*, as the model only specifies the speaker agent's beliefs about the listener's current mental state with respect to his/her contact, perception, understanding, acceptance, and agreement of what is being uttered. Computationally, this is modelled as sequential probabilistic inferences over a dynamic Bayesian network. This also allows the agent to quantify its uncertainty about the listener's mental state and to actively elicit feedback by means of dedicated cues (e.g. gazing at the listener while pausing speech) or even explicit queries. Note that this system does not employ the guiding principles of *hierarchical prediction-based processing* and *resonance-based behaviour perception and generation*.

In evaluation studies with human users [66,67], the attentive speaker agent was compared with agents that either do not adapt their social behaviour to their beliefs about listeners' needs (lower bound) or do it incessantly and explicitly ask about it (upper bound). That is, the first condition decouples the social loop from the cognitive one, while the second couples them excessively but in non-adaptive, socially non-resonant ways. We analysed first whether human interlocutors were willing to provide feedback to artificial agents at all. Results show that they only did so if the agent was actually attentive to their feedback and responded to it by adapting its behaviour (figure 3*b*). Second, we investigated whether participants perceived the agents to be different and whether they noted the attentiveness and adaptivity of the attentive speaker agent and were

(*a*)



(*b*)

model	steps	success	run-time	time per step
HAICA	60.79	0.89	**1.32**	**0.02**
BD	**36.52**	**0.92**	2111.55	55.87

Figure 4. (*a*) Outline of how the HAICA model for the green 'agent cook' perceives and acts on the *Overcooked* kitchen environment (including mentalizing over the blue 'agent cook'). (*b*) Results of simulations comparing the performance of HAICA, with optimal susceptibility parameters, with a full Bayesian reasoning and planning approach (BD) [70]. Comparison is done with regard to average numbers for required steps, joint task success rate, total run-time and time to plan one step ahead (best marked as bold). While BD needs about half as many steps as HAICA, its computational costs are three orders of magnitude higher and prevent using it in interactive settings. (Online version in colour.)

aware of the collaborative effort that it made to the interaction. Results demonstrate that participants noticed that both the attentive speaker agent as well as the upper bound control agent were attentive and adaptive, but only the attentive speaker agent was perceived as having a desire to be understood as well as being more helpful in resolving difficulties in participants' understanding.

(c) Collaborative situated problem-solving

As a final example, we turn to the question how socially enactive cognition could enable fast and robust task collaboration between agents. Classical work in AI on multi-agent coordination focuses on finding optimal solutions through planning, prior to the actual interaction and from the point of view of either a centralized, omniscient control unit [68] or a decentralized one through dedicated forms of communication between the agents. In recent work [69] we asked whether 'satisficing' (in contrast to optimal) collaboration can emerge without planning, only from on-the-fly coordination between socially enactive cognitive agents that rely on situated, prediction-based task behaviour that is affected by their mentalizing about other agents' intentions. To that end, a model of 'hierarchical active inference for collaborative agents' (HAICA; see figure 4, left) was developed. It continuously infers hypotheses about the mental states of another agent and integrates them into own prediction-based goal and intention formation through a mechanism we call *belief resonance*. Figure 4*a* shows an outline of the HAICA model, here for the agent with the green hat, with the mentalizing component in blue and the predictive processing hierarchy in green.

Inferring beliefs and goals of the collaboration partner is based on a Bayesian Theory of Mind approach (BToM) [55,71], which usually requires the evaluation of all possible actions and expected rewards (as a likelihood). HAICA's belief resonance uses a simplified approximation of that likelihood which assumes that the other agent would behave similar to oneself, given it finds itself in the same situation with the same intentions and goals at its disposal. This creates a set of beliefs that are integrated into the agent's intention or goal posteriors. The underlying update mechanism is modulated according to a *susceptibility parameter* that controls how strongly the mental states of another agent impacts one's own predicted future behaviour. Note that, aside from mentalizing about the other agent, HAICA does not involve a separation of subjective goals or intentions in its selection of actions. As a result, the model does not afford any mechanisms for planning ahead or supporting another agent with complementary actions.

Despite this quite simplifying (yet in many situations satisficing) similarity assumption, simulation studies yield highly informative and competitive results. Two agents equipped with HAICA were situated in the *Overcooked* game scenario, in which they have to prepare frequently ordered meals given limited resources. Agents were not equipped with simple-to-follow recipes. Rather, the hierarchical model represents a number of intentions of observations and actions that are chosen specifically for the *Overcooked* game scenario, along with a set of possible goals (in the form of meal orders) that influence the likelihood of inferring specific intentions. Agents are then run to detect and act upon afforded intentions, e.g. in response to the observed behaviour of the other agent or the observed availability of a resource that is likely to be needed for the given meal order.

The simulations show that two of those socially enactive cognitive agents, when paired together in the environment, are able to collaboratively succeed with this task in many situations and in a very resource-efficient manner. Achieved team performance is comparable to state of the art deep reinforcement learning models [70] and, although not always being optimal, is by orders of magnitude more computationally efficient than full Bayesian planning (figure 4*b*; see Pöppel *et al.* [69] for details). These results are due to an emergent coordination as the agents dynamically and bi-directionally adapt to each other as well as to the shared environment they simultaneously alter, yet without any explicit coordination through communication or global planning. This work is an example of multi-agent coordination that is implicit and highly situated, thanks to principles of socially enactive cognitive systems: The agents perceive the affordances that their environment provides and integrate this with social information about the other agents, following the principle of *resonance-based behaviour perception and*

generation. The underlying belief resonance mechanism affords a form of satisficing decision-making that does not require complex social reasoning or planning, but performs goal and action selection by way of minimizing predictive uncertainty (as described above) and thus is in line with the principles of *hierarchical prediction-based processing* and *differential and good-enough mentalizing*.

4. Conclusion

In order for intelligent systems to be able to engage in extended, meaningful interactions with humans we propose to work towards 'artificial social enactivism', according to which a social understanding and a social interaction are actively co-constructed by human and artificial interactants. In contrast to relying on complex models, long-term planning or explicit interaction protocols, we emphasize the view that the dynamics of the intra-agent socio-cognitive processing loop and the inter-agent social interaction loop evolve in parallel and in a bi-directionally constitutive fashion. We conjecture that humans prefer or even rely on this quality of social interaction, and that artificial systems consequently should enable this in order to be more human-compatible.

It is not clear yet what the best construction and computational modelling principles for socially enactive systems are in detail. However, recent research on embodied face-to-face interaction, enactive cognition, or social neuro-cognitive processes like the interactive brain hypothesis allows us to distil some of them that we deem to be implementable in artificial systems. First, resonance-based behaviour perception and action are needed to enable fast alignment and coordination phenomena in social environments. Second, hierarchical prediction-based processing is put forward to integrate perception–reasoning–action over time and over different levels of mental state abstraction. To that end, the FEP can be adopted to form predictions and on-demand inferential processes while staying compatible with theory-bound top-down mentalizing. Third, differential and good-enough mentalizing is required for a system to be able to differentiate between subjective mental perspectives in an efficient manner, while allowing bootstrapping mentalizing processes, possibly through abductive biases (e.g. similarity-based we-beliefs or a readiness to interact). This rejects the radical enactive stance of getting rid of mental representations. Instead we demand systems to make use of complex mental representations only where required. Finally, we endorse the principle of incremental processing to ensure and increase a system's responsiveness. This is thought of here as a means of incrementally coordinating the perception and production of socio-communicative behaviour in multi-layered coordination loops. The implemented examples that we have described demonstrate different features of artificial socially enactive systems by employing suitable combinations of the described key principles, hence paving the way for artificial social minds that ultimately enable truly human-compatible interactions.

Data accessibility. This article has no additional data.

Authors' contributions. S.Ka. and S.Ko. contributed equally to the preparation of the paper. S.Ka.: writing—original draft, writing—review and editing; S.Ko.: conceptualization, writing—original draft, writing—review and editing.

Both authors gave final approval for publication and agreed to be held accountable for the work performed herein.

Conflict of interest declaration. Both authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest in the subject matter or materials discussed in this manuscript.

Funding. This research was partially supported by the German Research Agency (DFG) in the priority programme 'The Active Self' and the Cluster of Excellence 'Cognitive Interaction Technology' (CITEC).

Endnote

[1]A video of this multi-agent belief coordination is available online: http://purl.com/skahl/interaction_video.

References

1. Chakraborti T, Sreedharan S, Kambhampati S. 2020 The emerging landscape of explainable automated planning & decision making. In *Proc. 29th Int. Joint Conf. Artificial Intelligence, IJCAI-20, Yokohama, Japan, 7–15 January 2021* (ed. C Bessiere), pp. 4803–4811. International Joint Conferences on Artificial Intelligence Organization.
2. Kambhampati S. 2019 Challenges of human-aware AI systems. *arXiv*, 1910.07089. (doi:10.48550/arXiv.1910.07089)
3. Pentland A. 2005 Socially aware, computation and communication. *Computer* **38**, 33–40. (doi:10.1109/MC.2005.104)
4. Kruse T, Pandey AK, Alami R, Kirsch A. 2013 Human-aware robot navigation: a survey. *Rob. Auton. Syst.* **61**, 1726–1743. (doi:10.1016/j.robot.2013.05.007)
5. van der Schyff D, Schiavio A, Walton A, Velardo V, Chemero A. 2018 Musical creativity and the embodied mind: exploring the possibilities of 4E cognition and dynamical systems theory. *Music Sci.* **1**. (doi:10.1177/2059204318792319)
6. Schilbach L, Timmermans B, Reddy V, Costall A, Bente G, Schlicht T, Vogeley K. 2013 Toward a second-person neuroscience. *Behav. Brain Sci.* **36**, 393–414. (doi:10.1017/S0140525X12000660)
7. Przyrembel M, Smallwood J, Pauen M, Singer T. 2012 Illuminating the dark matter of social neuroscience: considering the problem of social interaction from philosophical, psychological, and neuroscientific perspectives. *Front. Hum. Neurosci.* **6**, 190. (doi:10.3389/fnhum.2012.00190)
8. Premack D, Woodruff G. 1978 Does the chimpanzee have a theory of mind? *Behav. Brain Sci.* **1**, 515–526. (doi:10.1017/S0140525X00076512)
9. Churchland PM. 1981 Eliminative materialism and the propositional attitudes. *J. Philos.* **78**, 67–90.
10. Goldman AI. 1989 Interpretation psychologized. *Mind Lang.* **4**, 161–185. (doi:10.1111/j.1468-0017.1989.tb00249.x)
11. Becchio C, Sartori L, Castiello U. 2010 Toward you: the social side of actions. *Curr. Direct. Psychol. Sci.* **19**, 183–188. (doi:10.1177/0963721410370131)
12. Di Paolo EA, De Jaegher H. 2012 The interactive brain hypothesis. *Front. Hum. Neurosci.* **6**, 163. (doi:10.3389/fnhum.2012.00163)
13. Hurley S. 2008 The shared circuits model (SCM): how control, mirroring, and simulation can enable imitation, deliberation, and mindreading. *Behav. Brain Sci.* **31**, 1–22; discussion 22–58. (doi:10.1017/S0140525X07003123)
14. De Jaegher H, Di Paolo E. 2007 Participatory sense-making: an enactive approach to social cognition. *Phenomenol. Cogn. Sci.* **6**, 485–507. (doi:10.1007/s11097-007-9076-9)

15. Gallagher S. 2008 Direct perception in the intersubjective context. *Conscious Cogn.* **17**, 535–543. (doi:10.1016/j.concog.2008.03.003)
16. Shapiro L, Spaulding S. 2021 Embodied cognition. In *The Stanford encyclopedia of philosophy* (ed. EN Zalta). Stanford, CA: Metaphysics Research Lab, Stanford University.
17. Di Paolo EA. 2005 Autopoiesis, adaptivity, teleology, agency. *Phenomenol. Cogn. Sci.* **4**, 429–452. (doi:10.1007/s11097-005-9002-y)
18. Maturana HR, Varela FJ. 1980 *Autopoiesis and cognition: the realization of the living*. Dordrecht, The Netherlands: Springer Netherlands.
19. Bruner JS. 1978 *Toward a theory of instruction*, 8th edn. Cambridge, MA: Harvard University Press.
20. Vygotsky LS. 1978 *Mind in society: the development of higher psychological processes*. Cambridge, MA: Harvard University Press.
21. De Jaegher H, Di Paolo E, Adolphs R. 2016 What does the interactive brain hypothesis mean for social neuroscience? A dialogue. *Phil. Trans. R. Soc. B* **371**, 20150379. (doi:10.1098/rstb.2015.0379)
22. Di Paolo EA, De Jaegher H. 2015 Toward an embodied science of intersubjectivity: widening the scope of social understanding research. *Front. Psychol.* **6**, 234. (doi:10.3389/fpsyg.2015.00234)
23. Reddy V. 2003 On being the object of attention: implications for self–other consciousness. *Trends Cogn. Sci.* **7**, 397–402. (doi:10.1016/S1364-6613(03)00191-8)
24. Friston K. 2010 The free-energy principle: a unified brain theory? *Nat. Rev. Neurosci.* **11**, 127–138. (doi:10.1038/nrn2787)
25. Gallotti M, Frith CD. 2013 Social cognition in the we-mode. *Trends Cogn. Sci.* **17**, 160–165. (doi:10.1016/j.tics.2013.02.002)
26. Kahl S, Kopp S. 2015 Towards a model of the interplay of mentalizing and mirroring in embodied communication. In *Proc. EuroAsianPacific Jt Conf. Cognitive Science, 25–27 September, Torino, Italy* (eds G Airenti, BG Bara, G Sandini), pp. 300–305. Aachen, Germany: CEUR Workshop Proceedings.
27. Kahl S. 2020 *Social motorics - a predictive processing model for efficient embodied communication*. PhD thesis, University of Bielefeld, Bielefeld, Germany. (doi:10.4119/unibi/2945718)
28. Kopp S. 2010 Social resonance and embodied coordination in face-to-face conversation with artificial interlocutors. *Speech Commun.* **52**, 587–597. (doi:10.1016/j.specom.2010.02.007)
29. Wagner P, Malisz Z, Kopp S. 2014 Gesture and speech in interaction: an overview. *Speech Commun.* **57**, 209–232. (doi:10.1016/j.specom.2013.09.008)
30. Kopp S, Allwood J, Grammer K, Ahlsen E, Stocksmeier T. 2008 Modeling embodied feedback with virtual humans. In *Modeling communication with robots and virtual humans* (eds I Wachsmuth, G Knoblich), pp. 18–37. Berlin, Germany: Springer.
31. Giles H, Coupland N. 1991 *Language: contexts and consequences*. Milton Keynes, UK: Open University Press.
32. Clark HH. 1996 *Using language*. Cambridge, UK: Cambridge University Press.
33. Pickering MJ, Garrod S. 2013 An integrated theory of language production and comprehension. *Behav. Brain Sci.* **36**, 329–347. (doi:10.1017/S0140525X12001495)
34. Lakin JL, Jefferis VE, Cheng CM, Chartrand TL. 2003 The chameleon effect as social glue: evidence for the evolutionary significance of nonconscious mimicry. *J. Nonverbal Behav.* **27**, 145–162. (doi:10.1023/A:1025389814290)
35. Miles LK, Nind LK, Macrae CN. 2009 The rhythm of rapport: interpersonal synchrony and social perception. *J. Exp. Social Psychol.* **45**, 585–589. (doi:10.1016/j.jesp.2009.02.002)
36. Wang I, Ruiz J. 2021 Examining the use of nonverbal communication in virtual agents. *Int. J. Hum. Comput. Interact.* **37**, 1648–1673. (doi:10.1080/10447318.2021.1898851)
37. Brass M, Bekkering H, Prinz W. 2001 Movement observation affects movement execution in a simple response task. *Acta Psychol.* **106**, 3–22. (doi:10.1016/S0001-6918(00)00024-X)
38. Molenberghs P, Cunnington R, Mattingley JB. 2012 Brain regions with mirror properties: a meta-analysis of 125 human fMRI studies. *Neurosci. Biobehav. Rev.* **36**, 341–349. (doi:10.1016/j.neubiorev.2011.07.004)
39. Cook R, Bird G, Catmur C, Press C, Heyes C. 2014 Mirror neurons: from origin to function. *Behav. Brain Sci.* **37**, 177–192. (doi:10.1017/S0140525X13000903)
40. Heyes C. 2009 Where do mirror neurons come from? *Neurosci. Biobehav. Rev.* **34**, 575–583. (doi:10.1016/j.neubiorev.2009.11.007)
41. Prinz W. 1990 A common coding approach to perception and action. In *Relationships between perception and action* (eds O Neumann, W Prinz), pp. 167–201. Berlin, Germany: Springer
42. Hommel B, Müsseler J, Aschersleben G, Prinz W. 2001 The theory of event coding (TEC): a framework for perception and action planning. *Behav. Brain Sci.* **24**, 849–878; discussion 878–937. (doi:10.1017/S0140525X01000103)
43. Dumas G, de Guzman GC, Tognoli E, Kelso JAS. 2014 The human dynamic clamp as a paradigm for social interaction. *Proc. Natl Acad. Sci. USA* **111**, E3726–E3734. (doi:10.1073/pnas.1407486111)
44. Dumas G, Moreau Q, Tognoli E, Kelso JAS. 2020 The human dynamic clamp reveals the fronto-parietal network linking real-time social coordination and cognition. *Cereb. Cortex* **30**, 3271–3285. (doi:10.1093/cercor/bhz308)
45. Grossberg S. 1980 *How does a brain build a cognitive code?* Dordrecht, The Netherlands: Springer.
46. Kilner JM, Friston KJ, Frith CD. 2007 Predictive coding: an account of the mirror neuron system. *Cogn. Process* **8**, 159–166. (doi:10.1007/s10339-007-0170-2)
47. Clark A. 2013 Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behav. Brain Sci.* **36**, 181–204. (doi:10.1017/S0140525X12000477)
48. de Bruin L, Strijbos D. 2015 Direct social perception, mindreading and Bayesian predictive coding. *Conscious Cogn.* **36**, 565–570. (doi:10.1016/j.concog.2015.04.014)
49. Hohwy J. 2016 The self-evidencing brain. *Noûs* **50**, 259–285. (doi:10.1111/nous.12062)
50. Adams RA, Shipp S, Friston KJ. 2012 Predictions not commands: active inference in the motor system. *Brain Struct. Funct.* **218**, 611–643. (doi:10.1007/s00429-012-0475-5)
51. Friston KJ, Daunizeau J, Kilner J, Kiebel SJ. 2010 Action and behavior: a free-energy formulation. *Biol. Cybern* **102**, 227–260. (doi:10.1007/s00422-010-0364-z)
52. Cisek P. 2007 Cortical mechanisms of action selection: the affordance competition hypothesis. *Phil. Trans. R. Soc. B* **362**, 1585–1599. (doi:10.1098/rstb.2007.2054)
53. Matheson C, Poesio M, Traum D. 2000 Modelling grounding and discourse obligations using update rules. In *Proc.1st Mtng North American Chapter of the Association for Computational Linguistics, 29 April–4 May 2000, Seattle, WA* (ed. J Wiebe), p. 8. Stroudsburg, PA: Association for Computational Linguistics.
54. Larsson S, Traum DR. 2000 Information state and dialogue management in the TRINDI dialogue move engine toolkit. *Nat. Lang. Eng.* **6**, 323–340. (doi:10.1017/S1351324900002539)
55. Pöppel J, Kopp S. 2018 Satisficing mentalizing: Bayesian models of theory of mind reasoning in scenarios with different uncertainties. In *Proc.17th Int. Conf. Autonomous Agents and Multiagent System, 10–15 July 2018, Stockholm, Sweden*. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems.
56. Healey PGT, Mills GJ, Eshghi A, Howes C. 2018 Running repairs: coordinating meaning in dialogue. *Top. Cogn. Sci.* **10**, 367–388. (doi:10.1111/tops.12336)
57. Friston KJ, Frith CD. 2015 A duet for one. *Conscious Cogn.* **36**, 390–405. (doi:10.1016/j.concog.2014.12.003)
58. Vasil J, Badcock PB, Constant A, Friston K, Ramstead MJD. 2020 A world unto itself: human communication as active inference. *Front. Psychol.* **11**, 417. (doi:10.3389/fpsyg.2020.00417)
59. Veissière SPL, Constant A, Ramstead MJD, Friston KJ, Kirmayer LJ. 2019 Thinking through other minds: a variational approach to cognition and culture. *Behav. Brain Sci.* **43**, e90. (doi:10.1017/S0140525X19001213)
60. Kopp S, van Welbergen H, Yaghoubzadeh R, Buschmeier H. 2013 An architecture for fluid real-time conversational agents: integrating incremental output generation and input processing. *J. Multimodal User Interfaces* **8**, 97–108. (doi:10.1007/s12193-013-0130-3)
61. xKahl S, Kopp S. 2018 A predictive processing model of perception and action for self-other distinction. *Front. Psychol.* **9**, 2421. (doi:10.3389/fpsyg.2018.02421)

62. Frith CD. 2012 The role of metacognition in human social interactions. *Phil. Trans. R. Soc. B* **367**, 2213–2223. (doi:10.1098/rstb.2012.0123)
63. Fernandez-Duque D, Baird JA, Posner MI. 2000 Executive attention and metacognitive regulation. *Conscious Cogn.* **9**, 288–307. (doi:10.1006/ccog.2000.0447)
64. Bavelas J, Gerwing J, Healing S. 2017 Doing mutual understanding. Calibrating with micro-sequences in face-to-face dialogue. *J. Pragmat.* **121**, 91–112. (doi:10.1016/j.pragma.2017.09.006)
65. Allwood J, Nivre J, Ahlsén E. 1992 On the semantics and pragmatics of linguistic feedback. *J. Semant.* **9**, 1–26. (doi:10.1093/jos/9.1.1)
66. Buschmeier H, Kopp S. 2018 Communicative listener feedback in human–agent interaction: artificial speakers need to be attentive and adaptive. In *Proc.17th Int. Conf. Autonomous Agents and Multiagent Systems, 10–15 July 2018, Stockholm, Sweden* (eds M Dastani, G Sukthankar, E André, S Koenig), pp. 1213–1221. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems.
67. Buschmeier H. 2017 Attentive speaking: from listener feedback to interactive adaptation. Dissertation, Bielefeld University, Bielefeld, Germany.
68. Torreño A, Onaindia E, Komenda A, Štolba M. 2018 Cooperative multi-agent planning: a survey. *ACM Comput. Surv.* **50**, 1–32. (doi:10.1145/3128584)
69. Pöppel J, Kahl S, Kopp S. 2021 Resonating minds–emergent collaboration through hierarchical active inference. *Cogn. Comput.* **14**, 581–601. (doi:10.1007/s12559-021-09960-4)
70. Wu SA, Wang RE, Evans JA, Tenenbaum JB, Parkes DC, Kleiman-Weiner M. 2021 Too many cooks: Bayesian inference for coordinating multi-agent collaboration. *Top. Cogn. Sci.* **13**, 414–432. (doi:10.1111/tops.12525)
71. Baker CL, Saxe R, Tenenbaum JB. 2011 Bayesian theory of mind: modeling joint belief-desire attribution. In *Proc. Annu. Mtng Cognitive Science Society, 20–23 July 2011, Boston, MA* (eds L Carlson, C Hoelscher, TF Shipley), pp. 2469–2474. Austin, TX: Cognitive Science Society.