Phil. Trans. R. Soc. B. article template

Paper forthcoming, please cite published version

PHILOSOPHICAL TRANSACTIONS B

Phil. Trans. R. Soc. B.
doi:10.1098/not yet assigned

What does the Interactive Brain Hypothesis mean for Social Neuroscience? A dialogue.

Hanne De Jaegher[1,2], Ezequiel Di Paolo[3,1,2], and Ralph Adolphs[4,5,6]

1 Department of Logic and Philosophy of Science, IAS-Research Centre for Life, Mind, and Society, University of the Basque Country, San Sebastián, Spain.
2 Department of Informatics, Centre for Computational Neuroscience and Robotics, and Centre for Research in Cognitive Science, University of Sussex, Brighton, UK.
3 Ikerbasque, Basque Foundation for Science, Bilbao, Spain.
4 Computation and Neural Systems, California Institute of Technology, Pasadena, CA 91125;
5 Division of Humanities and Social Sciences, California Institute of Technology, Pasadena, CA 91125
6 Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA 91125

Keywords: Social neuroscience, interactive brain hypothesis, social interaction, social cognition, causality, participatory sense-making

Summary

A recent framework inspired by phenomenological philosophy, dynamical systems theory, embodied cognition, and robotics has proposed the Interactive Brain Hypothesis (IBH). Whereas mainstream social neuroscience views social cognition as arising solely from events in the brain, the IBH argues that social cognition requires in addition causal relations between the brain and the social environment. We discuss, in turn, the foundational claims for the IBH in its strongest form; classical views of cognition that can be raised against the IBH; a defense of the IBH in light of these arguments; and a response to this. Our goal is to initiate a dialogue between cognitive neuroscience and enactive views of social cognition. We conclude by suggesting some new directions and emphases that social neuroscience might take.

1. Introduction

In the context of recent advances in social neuroscience, particularly the availability of methods for investigating brain activity in complex situations, including live interaction, novel research questions emerge. Do brains in interaction function the same way as in non-interactive situations? Are social interactions simply more complex scenarios involving more dynamical kinds of processing, but not essentially different from non-interactive cases? Or do live social engagements between people engender novel phenomena, which prompt us to reconsider brain function?

A recent proposal strongly vouches for the last option: Brains work differently in social interactive situations. And moreover, the dynamics of the interaction itself play important roles in cognitive function [1].

We investigate some implications of this view by raising critical questions. This will take the form of a dialogue between the authors of this paper – not to settle the issues or to iron out wider conceptual disagreements once and for all, but to progress, if not to a final common ground, then hopefully to some useful inroads into it.

2. The IBH and how social neuroscientists should view it (HDJ and EDP)

Two of us (De Jaegher and Di Paolo) argue that embodiment and interaction are partly, but fundamentally, constitutive of social cognition. This view is captured by the Interactive Brain Hypothesis (IBH): "The IBH ... proposes that social interaction processes play enabling and constitutive roles in the development and in the ongoing operation of brain mechanisms involved in social cognition, *whether the person is engaged in an interactive situation or not*" [1,p.2]. We use here the terminology introduced in [2]: an enabling factor is causally necessary for a phenomenon to occur, while a constitutive factor is part of what makes the phenomenon what it is. While a hypothesis rather than a scientific claim, the IBH stimulates a novel perspective on how social neuroscientists should construe information processing that generates social behavior. The unit of analysis is no longer delimited to the brain, but broadened to include aspects of the social environment with which the brain interacts or has interacted.

There is a range of hypotheses about the role that embodiment and social context play in social cognition. The weakest claim is that social interaction is methodologically useful in social neuroscience, since it provides ecological validity, and engages research participants. Another weak claim is that social interaction needs to be considered as providing important contextual modulation of the social brain. To our knowledge, nobody disagrees with these claims. They are now actively pursued in social neuroscience, as evidenced in the other contributions in this issue; we do not further treat them here.

A stronger claim is that social interaction facilitates particular kinds of brain processes: that is, there is a strong enabling role for social interaction [2]. For instance, it is well known that development in the absence of social interaction (severe social deprivation, in humans or other mammals) results in a highly abnormal brain with

*Author for correspondence (h.de.jaegher@gmail.com).
†Present address: Department of Logic and Philosophy of Science, University of the Basque Country, Av. De Tolosa 70, 20018 San Sebastián SPAIN

highly abnormal cognition [3]. This fact also suggests important constraints on the design of artificial cognitive systems. To our knowledge, this claim is uncontentious as well. We do not discuss this enabling aspect of social cognition here either.

The hypothesis we discuss here concerns an occurrent instance of social cognition. We claim that a normal adult human brain in isolation is insufficient for a typical instance of social cognition. Of course, we acknowledge that the brain plays a large, and likely major, role in social cognition.[1] But processes occurring in that brain at the time of an instance of social cognition are, in the typical case, not fully constitutive of social cognition: additional events are required as well. Those additional events involve *relations* between the brain and (parts of) the rest of the world. It is important to note that we are not claiming that events external to the brain, in isolation, make a determining contribution. It is the relations between such external events, the body, and the brain that matter, or rather, it is only within these relations, which are not merely contextual, that we can make sense of brain function during most social cognition. In such cases, there is no factoring out of such extra-brain elements without removing at the same time something essential to social cognition as such.

What are those instances of social cognition where we expect extra-neural relational patterns to play a constitutive role? Certainly those involving direct interaction with others. But also those instances involving the presence of others to various degrees (physical, virtual, etc.) and which predispose us to engage interactively, even if we do not or cannot actualise such dispositions.

FIGURE 1 HERE

Consider an example where, we argue, social interaction plays a constitutive role in the performance of a social task. This is the perceptual crossing experiment by Auvray et al. [4] (Figure 1). Two blindfolded participants are told to freely move a sensor (a computer mouse) along a shared virtual line. In this virtual 'world', each participant can encounter two different kinds of objects, one fixed in space, and two moving objects. One of the moving objects corresponds to the scanning sensor of the other participant, and the other is attached to this sensor at a fixed distance, like a "shadow". In terms of their trajectories, the moving objects are indistinguishable. Whenever her sensor encounters an object on this line, the participant receives an on/off tactile stimulus – a tap on the finger, which is the same for each kind of object. This situation is symmetrical for both participants. Notice that when a participant's sensor encounters the shadow of the other participant, only the first participant will receive a tactile stimulus. When the two sensors meet, both participants receive a stimulus simultaneously.

Participants are instructed to click the mouse button whenever they judge they are in contact with the other participant. As a result, statistically, mouse clicks tend to concentrate on each other's sensors (65.9% of clicks) and not on the identically moving shadow objects (23%). This means that participants can find each other by "perceptually crossing" their scanning activities. However, the authors find that the probability of clicking following a stimulus is approximately the same whether this stimulation comes from the other's sensor or its shadow. Auvray et al. explain this result in terms of the collective dynamics of the interactive configuration. Each participant attempts to re-scan many times an object that is apparently moving, by making back and forth mouse movements. When this scanning involves an encounter between both sensors, the participants tend to continue this activity for a long time. In contrast, when one of them is scanning the other's shadow, which is moving independently of this scanning movement, this 'encounter' is short-lived. This means that the dyadic system is organised such that sensor–sensor encounters are more frequent than sensor–shadow encounters, which explains why participants can "find each other" in spite of the sensory ambiguity (see [2]). For this reason, we consider that the interaction process here plays a constitutive role in how the social task is organised and performed.

As this and other examples show, we cannot always assume that social interactions are mere inputs to cognitive systems, but rather, in general, we need to consider as relevant *both* individual *and* interactive

---

[1] This is not to say that we think that understanding the brain is sufficient for understanding cognition. From an enactive perspective, what matters is the embodied subject in relation to her world or meaningful environment. Interestingly, the same kind of argument in favour of the IBH presented here in the context of social neuroscience could be made in the context of many approaches to embodied cognition that remain methodologically individualistic. For these, the body is crucial for the mind, meaning the individual body and not its engagement in social interactions. From the enactive perspective, in contrast, both body and social engagement are primordial [1,2].

mechanisms from the start [2]. The IBH proposes that social interactions are at the basis of social skills more often than usually assumed. Therefore, the point of the IBH is to provide a research guide to specify which social events and social relations, as well as what kind of brain activity, matter and how, to particular instances of social cognition. As methodologies for investigating brain activity during free interactions continue to develop, it is necessary to theorise about these questions, especially as we approach everyday situations involving emergent collective patterns, jointly authored actions, and multiple brains and bodies in coordination. With the IBH we question the tacit assumption that the best way to approach this challenge is to try to understand brain function in isolation by assigning to all extra-neural variation a role exclusively as inputs or outputs. Instead, we propose to leave open to investigation the conditions under which this assumption may be valid as a limiting case.

3. Arguments against the IBH (RA)

The other one of us (Adolphs) disagrees with the above view because it seems to disregard a natural partition to causal interactions in the world. That is, there is a much more direct and dense set of causal interactions internal to the brain, than between brain and external environment. Disregarding this fact leads to a concern that the IBH renders unclear the specific role of social neuroscience (as opposed to social psychology or sociology or behavioral studies of crowd behavior) in explaining social behavior.

In order to understand cognition, we need to partition cognitive systems. In particular, we need to partition them into those parts that should be analyzed as inputs to the system, those that are the outputs from the system, and those that are actually implementing the cognition. We do the same thing with computers running programs: there is a causal interaction with the world that can be treated as input, there is processing internal to the computer, and there is causal interaction with the world that can be treated as output.

The IBH follows the more dynamical systems view that much of situated cognition has adopted [5,6], and claims that this partitioning does not reflect how cognition actually takes place in the world. Unlike the classical computer metaphor, there is a more or less continuous stream of causal interaction between brain and world, and, in the case of the social world, "outputs" from a brain influence "inputs" (i.e., one person influences another) in such a tightly coupled way that it becomes impossible to distinguish input from output. Instead, say advocates of the IBH, one should treat the whole system (two or more people interacting) as a single, dynamically coupled system. Cognition is constituted by the events in my brain, the events in the other person's brain, and the causal relations between them: the whole system matters.

I am sympathetic with what motivates the IBH, and of course I agree that the classical computer metaphor is inadequate. Indeed, modern cognitive neuroscience acknowledges that much of cognition is "active": we continuously move our eyes to redirect visual input [7], we continuously shift attention to redirect what information is processed, we continuously interact with the environment especially in the case of a social encounter [8,9]. Current enthusiasm about Bayesian or predictive coding approaches [10,11] reflects this acknowledgment. But I am confused about how and why one would need to adopt the IBH, as opposed to one of its weaker forms, to incorporate these facts. To make my confusion more transparent, consider three properties of a person that we might want to understand: their observable behavior, their cognition, and their conscious experience. Let's consider these in light of an experiment that IBH advocates have mustered: the experiment by Auvray et al. [4], discussed above (cf. Figure 1). In this study, participants find each other's sensors, even though they themselves appear unaware of whether they are finding a sensor or a shadow. One could quibble about various aspects of this example as a good example of the IBH in action (it is not particularly "ecologically valid"; the fact that the subjects cannot explicitly distinguish sensor from shadow does not show that their brains are not representing this distinction, just unconsciously; etc.), but let us take it as an example nonetheless. Now the question is: what exactly does this experiment show? It shows that coupled causal interactions between two people are required to explain something – what is that? As far as I can tell, it is only a certain aspect of behavior. Yes, the behavior of the system cannot be explained only by events in individual brains. Me riding a bicycle also cannot be explained only by events in my brain. Much of our behavior comes about through complex causal interactions between our brains and the world, and social behavior is no exception.

Now, to see the limits of this example, ask yourself what the answer would be with respect to conscious experience. Is the coupled system of two people interacting supposed to be aware of the distinction between sensors and shadows? Surely not. One good reason is that whatever it is about the system that is generating the behavior of the system under consideration here seems far too meager an example of processing to count as cognition. The two people's brains in the experiment are each processing information so as to generate cognition and conscious experience. The entire system generates a unique behavior, but that is it. There is not

in addition any kind of collective "cognition" generated for the same reason that there is not in addition any kind of collective consciousness generated (intuitions here may of course diverge; see [12-14]). The reason is that the causal interactions at the systems level that explain the behavior are far too thin to constitute cognition. Cognition requires an extraordinarily dense and complex set of causal interactions that are part of an extensive processing architecture. Whatever exactly one's view on what cognition is, it is far more than a reflex, far more than a fixed action pattern, and instead is a highly inferential, context-dependent, and flexible form of information processing. So far as we know, only the brains of certain animals can generate examples of it. The causal interactions between those brains and the rest of the world are simply too "thin".

This brings me to my core objection against the IBH, as a hypothesis about cognition: in widening the causal base of cognition, it negates a distinction that is critical to understand cognition, the distinction between those causal events internal to the brain, and those constituting a brain's relations with the rest of the world. This distinction is huge. A brain's 80 billion or so neurons, or a much larger number of compartments of those neurons (opinions vary on what to consider the basic processing units in the brain) all causally interact, at multiple time scales. A single cortical neuron gets input from perhaps 10,000 other cells and participates in networks at local and global scales. Needless to say, we do not understand exactly how information is processed in the brain, but clearly it depends on very dense, very complex, sets of biological causal interactions between networks of cells in the brain. By comparison, the path of causal inputs to the brain (or outputs from it to the world) is extremely sparse. There are only a million axons from the eye going into the brain. There are many more axons between processing stages deeper in the brain (indeed, there are more axons from higher-level brain regions back down to lower-level regions, such as the "feedback" from visual cortex to thalamus, than in the opposite "feedforward" direction). In the auditory system, there are only about 3000 cells that transduce sound in each ear. Yet from this, through considerably more complex processing internal to the brain, we can hear music. Inside the brain there is cognition and conscious experience. Outside the brain there are causal relations, and indeed some of those causal relations can be fairly complex and reciprocal. But they are not part of the brain's computations and they are not also constitutive of cognition.

Indeed, cognition does not require concurrent causal interaction with the world at all: we can think, calculate the product of two numbers, and generate images with our eyes closed in a quiet room, or while dreaming. Moreover, a lot of such internal cognition is social: we think and dream about other people all the time. All of the occurrent causal events that constitute such examples of cognition must be limited to what happens inside our skulls (or perhaps also our bodies). The IBH deals with this problem by including in its substrate for cognition not only those causal relations occurring between brain and world at the time of the cognition, but also relations between brain and world that happened in the past. This has always struck me as a rather desperate move that brings us back to how we began this section. Yes, of course, cognition depends on the history of causal interactions with the world. Had my causal history been very different, my cognition would also be very different. But the reason for this difference should be apparent: the only mechanism by which my cognition could be changed in light of a different causal history is through the brain. Change my causal history, you change the brain and hence cognition. All this shows is that causal history is one particular kind of "input" to the brain over time (albeit one that might ultimately ground what it is that the brain's representations are about; see concluding section).

It may well be that "causal density" as I have described it above is just a proxy for another property that is more fundamental to cognition. Perhaps it is computational complexity. Perhaps it is something that requires much more clarification, like "ownership" for a person. Perhaps it is something like "manipulability" which could ground our concept of causation; one could imagine "manipulability" as experimental manipulability by us, or as biological manipulability in terms of what is accessible to evolution or development. Much more debate will be needed to develop arguments for any of these, but for present purposes "causal density" serves as a simple and intuitive metric.

There is no question that there are collective social phenomena that emerge from causal relations between multiple people and their shared environment. Group behaviors, politics, and the stock market are all examples of this. Each has its proper domain of study required to explain phenomena that emerge at those macroscopic levels. Disciplines like political psychology and economics tackle that. None of these truisms, however, challenge the neuroscience of social cognition: the proper domain of study to explain social cognition is the individual brain. The social neuroscientist does not also need to be studying the stock market. Even if the stock market were a cognitive system (unlikely in my view), then this still does not undermine the study of individual brains to understand social cognition. If eventually we engineer a computer so advanced that it has cognition, we would not also need to understand the cognition happening inside the brains of the people who built that computer. The reasons for the distinctions in all these examples are the same: they are just separate systems. Perhaps there is stock market cognition, AI, and human cognition. I can study them individually, and add chimpanzee cognition and pet dog cognition. What is important is to partition the world into systems, the internal constituents of which interact in ways that do not require also knowing how they interact with the rest of the world. Our understanding of the world requires such partitioning, and the disciplines that have arisen to explain how the world works reflect those partitions. The cognitive holism that

the IBH envisions erases real distinctions and, if carried through all the way, would make understanding cognition intractable because it is everywhere.

4. In defence of the IBH (EDP and HDJ)

Let us consider the IBH at its most radical: the claim that the dynamics of social interaction play constitutive roles in social cognition. The developmental version seems less controversial, although its implications are not trivial (see e.g. [15]). In fact, for any environmental factor to developmentally shape the function of brain processes, it cannot be systematically just an informational input. To play an informational role strictly requires the stationary functional context of the system for which a signal serves as an input. Hence, the developmental IBH also necessitates the possibility of interaction dynamics playing more than just informational roles.

Turning to the constitutional version of the IBH, we first must stress what the claim is. We defend that the dynamical processes involved in social interactions, which implicate not just extra-neural processes, but also relational processes between participants (and their surroundings) *can* be a constitutive part of the processes of social cognition as they are enacted by the individual participants involved [1,2]. The strong version of the IBH simply hypothesizes that this *possibility* is in general a widespread *plausibility*. This can be criticized in two ways: the extension from possibility to plausibility is not empirically warranted, or the very claim of possibility is wrong. The criticisms of the previous section are centered on the second option. If this possibility claim is wrong, then the constitutive version of IBH falls with it and only the developmental version remains.

We discuss three aspects in support of the constitutive claim: 1) the non-decomposability of neural and extra-neural processes during interaction, 2) the functional role of interaction dynamics, 3) the irreducibility of interactive phenomena, such as meaning generated in social interaction and the co-authorship of interactive acts.

1) Entanglement
The brain internal "causal density" argument discussed above seems compelling only if we assume that brains are "nearly decomposable" systems [16] with respect to body and environment. Nearly decomposable systems interact with other systems without losing their functionality or altering significantly their internal causal relations. Considering the brain in this way means to treat its couplings with body and environment as inputs. There are solid arguments against the disposability of body and environment for normal brain function. Some are based on the abundant evidence of the entangled neural, body, and environmental dynamics in a wide range of cognitive performance [17]. A more conceptual argument is Thompson and Cosmelli's critique of the brain-in-the-vat thought experiment [18]. They argue that it is inconceivable for a brain to retain its functionality if separated from body and world.

We could assume that the causal support given by body and environment does not constrain neural function and so, at least functionally, the brain could be considered independent. But even in such a case, we cannot infer near-decomposability from the evidence of inner causal density alone. We must also demonstrate that inner processes are not dominated, shaped, or regulated in their function by external processes, i.e., that coupling with the world does not involve non-linear interactions across a significant range of timescales. In short, the inner complexity of the brain, which is of course undisputed, is not a deciding factor between the two interpretations discussed here: interactional processes as input vs. interactional processes as constitutive of social cognitive function.

Consider the evidence of the entanglement of brain and interaction dynamics observed in dual-scanning experiments [19]. According to Simon [20, p.204] a nearly decomposable system "[separates] the high-frequency dynamics of a hierarchy – involving the internal structure of the components – from the low frequency dynamics – involving interactions among components." But this precisely is not the case during inter-brain synchronization in live interactions. Using dual EEG scanning during an imitation task with interactors visibly moving their hands freely and allowing spontaneous synchrony and turn-taking, Dumas et al. [21] found interbrain phase synchronization in the alpha-mu (8–12 Hz), beta (13–30 Hz) and gamma (31–48 Hz) bands. How can social interaction affect neural oscillation phase in two distinct brains at frequencies more than one order of magnitude faster than the interactive movements?

Leaving aside the question of what role (if any) might be played by such cross-scale synchronization, the evidence suggests that interaction patterns produce an entanglement between the brains of the participants. Internally, the wave of influence across various temporal and spatial scales may travel from low to high frequencies via variations in neuronal excitability [22-25]. These top-down effects, evidenced also in arrhythmic cross frequency couplings [26], have been associated with different cognitive phenomena, notably

with the control of visual attention [27-29]. From here it is not a big leap to suggest that inter-brain synchronization at high frequencies [21,30,31] is due to high-to-low frequency integration and low-to-high frequency enslavement, with the difference that, instead of slow neural oscillations, the processes "at the top of the hierarchy" are the emergent rhythms of social interaction. This seems the simplest interpretation of the data, not the only possible one. But until disproven, it is not a bad idea to follow Occam's advice.

This interpretation is in line with calls to investigate the braided coordination of neural, behavioural and social processes [32,33]. It also coheres with cumulative evidence of the brain-body as an interaction-dominant system (the opposite of a nearly decomposable one), based on findings of correlations of neural and behavioural variability across a wide range of timescales [34,35]. Interaction-dominant systems are characterized by the causal inextricability of the various processes involved, as well as the unpredictability of the behaviour of the whole from knowledge of the parts in isolation. Evidence of interaction-dominance has also been found to involve extra-neural factors, e.g., in agent–tool systems [36] and during social interaction [37-40].

In view of this evidence, our suggested explanation of multiscale inter-brain synchronization engendered by emergent interaction patterns seems plausible. This allows us to make two points. The first, which is negative, is that this evidence casts doubt on the causal density argument against the IBH. Indeed, it would seem that at least under some conditions, brain, body, and interactive activity are under mutual causal influence, despite (or thanks to) the density of causal linkages in the brain. The second, neutral point raised by entanglement is that if social interaction can have such an influence on brain activity, then it is clearly possible that the interactive influence on brain dynamics during instances of social cognition is of a functional kind. To this positive possibility we turn next.

2) Functional roles for social interaction

Evidence of entanglement suggests that we should discard the view of interaction patterns as mere inputs to compartmentalized brain processes. But it does not yet say whether this more complex picture is sufficient to warrant the interpretation that interaction dynamics can be constitutive of the functional aspects of social cognition. What kind of cognitive "work" could be done by social interaction? This question cannot be answered in general terms. Each case will merit its own response. But at least in some cases we can provide a story. This is the importance of experiments like perceptual crossing, mentioned before [4]. In it the ecological situation is maximally simplified without eliminating a key factor: the free control of the social interaction dynamics by the participants. We do not think that this is an example of "just behaviour", if by this is meant that no sense-making is involved. It is a powerful exemplar that thanks to its simplicity can help us think differently about individual and interactive processes in more complex cases.

The perceptual crossing task is anything but simple. It is only *after* the performance has been explained that it appears so. In fact, described in strict computational terms it is a highly ambiguous, type-2 problem [41], i.e. a problem where stimuli must be actively discriminated spatially and qualitatively using only temporal and proprioceptive cues (all "objects" found in the virtual space produce the exact same on-off tactile stimulation). The task set to the participants is no less complex than typical discrimination tasks. In fact, it is untypically difficult, since the two moving objects that the participant can encounter (the other participant's sensor and shadow) move identically. Distinguishing them would require, from an individual perspective, not only a complex strategy for testing socially contingent reactions in these objects but also measuring these reactions in a highly ambiguous sensory space.

The fact that this computationally tough problem can be resolved with relative ease in the presence of interactive dynamics doesn't make this too meager an example of social cognition. That its difficulty deflates dramatically once we understand the collective dynamics is precisely the theoretically pregnant point of the experiment.

The type-2 regularities in the sensory signals that could help distinguish sensors from shadows are statistically invisible in the absence of a systematic sampling strategy. One way to solve the task is to implement a strategy that successfully transforms type-2 signals into type-1 data, i.e., into non-relational and unambiguous inputs [41]. A type-1 signal by itself contains enough information to determine the next course of action towards the resolution of the task. This route towards solving the task involves a biased sampling of the raw sensory streams, such that the task is rid of its ambiguities. Were this biased sampling to be implemented in the participants' brains, we would not hesitate to acknowledge that the processes involved are responsible for the core cognitive workload required to solve the problem. In other words, to *solve* the perceptual crossing task using this strategy *is* to find the way of biasing the sampling of sensory inputs so as to transform them from type-2 into type-1.

Now, this sampling bias is precisely what is achieved by the collective dynamics, i.e., by the interactive combination of individual strategies. As shown by Auvray, et al. [4], the interaction process biases the statistical presentation of sensory stimulus towards much more frequent encounters with the other

participant's sensor and not the shadow. Mutual scanning of sensors produces mutual sensory feedback and a permanence in the shared spatial region. This is more stable than one participant uni-directionally scanning the shadow of the other, who is unaware of this scanning and continues the search in other areas; thus the scanned shadow object quickly disappears. This is not done consciously by the participants but by the relation between their correlated movements. This cognitive work is neither given externally (in which case, we would be right in attributing the solution of the problem to a third party) nor is it generated internally within the participants' brains. It is produced by the self-organized collective dynamics in which they participate but whose properties do not correspond to individual properties of either agent on its own or to a linear aggregation of these. The task is transformed from type-2 to type 1 – it is *solved* – by the interaction process. There is no need for the participants' brains to represent the distinction between sensor and shadow at all to solve the task. The participants reap the benefits and deal with quasi-disambiguated, type-1 stimuli: "if it moves but stays nearby (repeated crossings), then click". If a process instantiates the solution to a cognitive problem it *constitutes* an instance of cognition. This is what social interaction does in perceptual crossing.

Further empirical confirmation that social interaction can play constitutive roles in social cognition is provided by a variation of the perceptual crossing experiment by Froese et al. [42]. This variation involves a more sophisticated social cognitive faculty, that of recognizing the other as an agent. The authors found that if they instructed the participants in a perceptual crossing task to cooperate as a team in finding each other, through several repeated interactions, the probability of clicking on the other's sensor grew to twice as much as that of clicking on the shadow object (in the original experiment these probabilities are approximately the same; the difference in absolute clicks is given by the interactively skewed probability of encountering each object). This means that participants develop a better way of "telling" if they are in contact with another agent, for instance, by using prototypical, co-authored regularities in the interaction patterns, which in turn would confirm the direct co-presence of the participants. Some pairs developed clear turn-taking patterns. As the authors say, these co-authored patterns turned "the *individual epistemic* task of agency detection into a *social pragmatic* task aimed at mutual coordination" [42, p. 4]. Since mutual recognition is a fundamental aspect of a wide range of cases of social cognition, its social constitution in as simple a situation as perceptual crossing is suggestive of an interactive sharing of socio-cognitive processes in other cases.

3) Irreducibility

The examples of entanglement and cognitive functionality evidenced in at least some cases of social interaction are indicative of phenomena that cannot be fully determined by what goes on in the individual participants' brains and bodies. But there is also an important sense in which the acts and meanings that are cognized about in social cognition are themselves part of emergent interactive phenomena and not simply a summation of individual attributes (such as moods, intentions, etc.). To cognize socially, in the enactive understanding of the term, is to skillfully engage in the multiple demands and possibilities of the social world, many of which are directly or indirectly emergent from social interactions. During interactive encounters, this skillful engagement does not in general necessitate tracking evidence that allows us to infer the mental states of others. Often such mental states do not directly impact on what is immediately required at the present moment or they are directly evident in the acts and responses of the others. Crucially, in such situations of interactive engagement, it is not individual cognizing and behaviour that sufficiently determines the relevant phenomena: both social acts and meanings are constituted socially and during the interactive encounter – think of a handshake, or the act of giving/receiving an object. The interactive constitution of social acts and meanings is a joint cognitive process that necessitates, but is under-determined by, individual cognition; the remainder of determination is given by the relational dynamics of the interactive encounter. We call this process participatory sense-making [43].

Consider escalation as a simple example of what we mean by irreducibility in the case of interactive phenomena. Typically, escalation involves an antagonistic pattern of interaction, sustained in time, increasing in intensity, and potentially spiralling out of control. Past conflictive interactions can predispose the onset of escalation even when the interaction partners do not individually intend to engage in an antagonistic exchange (see for instance, [44]). Sometimes escalation arises spontaneously as a result of interactive patterns. An example is given by Shergill et al. [45]. The interaction is quite minimal and involves participants applying a downward push with a finger on the other one's hand; an operation that is then repeated alternating roles. Participants are instructed to apply the same force as the perceived external force applied to them. As the turns alternate, the absolute amount of force escalates. The suggestion is that participants tend to underestimate the force they apply: self-generated force is perceived as weaker than externally generated force. Participants compensate by increasing force in the next round, resulting in escalation. Simple as this explanation is, it provides a good model for more general situations: escalation can originate unintentionally by a reciprocal configuration in which a perceived mismatch between one's own "moves" and those that we are subject to by external action.

The full explanation in this case combines individual and relational factors: a tendency to underestimate one's own force and the configuration of the alternating interaction pattern. Remove either factor and the

explanation fails. Moreover, the explanation does not involve any high-level awareness of the escalating pattern or deliberate intention to initiate escalation. Like the perceptual crossing situation, the onset of escalation just happens as part of the collective dynamics.

This simple case exemplifies how interactional dynamics are not fully under the control of the participants. There is no escalation module in the brain or an individual intention to escalate in general. It also shows that an important aspect of social meaning can relate to these emergent patterns. Escalation is often associated with the generation of negative affect, which undoubtedly relates to interactional history, but as we can see, can emerge as novel social meaning due to the interaction itself and not to any individual intentions. Similar processes where social meaning is generated by interactional patterns were already described by Gregory Bateson in terms of schismogenesis and feedback [46], and taken up in psychotherapeutic contexts (see e.g. [47-49]). The objects of social meaning are themselves interactively generated as well as apprehended.

Social interaction processes can be very hard to disentangle from individual brain and body dynamics. They can also play specific functional roles in the solution of a cognitive task. And they can give rise to objects, meanings, and actions that are irreducibly interactive. These complex realities in no way eliminate the possibility of scientific inquiry. On the contrary, in some cases they result in simpler explanations than those that are unduly constrained to be skull-bound, as we witness in the case of escalation and perceptual crossing. Far from making social cognition fuzzier and mysterious, the IBH in fact seeks to provide a more objective foundation, one that is more amenable to scientific observation and experimental manipulation.

5. Response (RA)

My co-authors are correct in taking me to reject the possibility, not just the plausibility, of the IHB in what I wrote in section 3. However, this depends on three concepts, whose relations were argued for only in the vaguest terms; let me say a bit more about them here in responding to the arguments of section 4. The three concepts are cognition, causation, and explanatory domain (for a discipline). Very roughly, my idea was that features of causation ("causal density") put constraints on cognition, and that this put constraints on what social neuroscience can study in order to understand cognition.

None of us has defined what we mean by "cognition", although I alluded to two other concepts as perhaps providing some reference: computation and consciousness. If cognition is taken to be processing that could at least potentially contribute to the contents of our conscious experience [50], then I found it implausible that the coupled system of the Auvray experiment had cognition. HDJ and EDP do not seem to have the same concept of "cognition", and instead theirs may rely more on how coordination (of behavior, of meaning) arises from social interactions. I am unclear on what my co-authors mean by "cognition"; but I am also unclear on what I myself mean. So I think this is one obvious way forward in our discussion: insofar as all of us are vague on what we mean by "cognition", it opens the way for a revised understanding of this concept that might reconcile our apparent differences.

How would one go about revising a concept of "cognition"? One place to begin would be by taking the term as relative to a discipline. This brings us to the topic of "explanatory domain". The points made in the previous section all argue that the study of the brain alone is insufficient to understand the kind of coupling we see in social interactions. I found the example of the Auvray experiment too detached from what happens in the brain, but HDJ and EDP argue it is not simple, not an atypical example, and not reducible to events in the brain and events outside the brain. In short, the suggestion of section 4 is that social neuroscience could gain more traction on how it uses the concept of "cognition" to explain behavior, if it incorporated relations with extra-neural events into its domain of study. This is an empirical suggestion: social neuroscientists should try to take this stance, and see how far they get with it. Will it be helpful in explaining human social behavior, or will it create complications if we widen the discipline of social neuroscience in this way? This seems like a reasonable practical position. If "cognition" is somewhat relativized to a discipline in this way, shifts in the explanatory domain of the discipline would result in corresponding shifts in the concept of "cognition".

The final issue concerns causation: I felt that this was much "denser" in the brain than between brain and environment, but the only metric I offered were sheer numbers of axons. HDJ and EDP argue that this is not the right metric, since even very small physical connections can result in profound influences. I think they are right. This then leaves me to retreat to something other than "causal density" as the distinguishing feature that delimits processing in the brain from processing involving events outside the brain. The only other good metric that comes to mind is something like "evolvability" or "manipulability by evolution". That is, there is a strong intuition that evolution can direct changes in cognition through changes in the brain, but not changes in the physical environment. Unfortunately this intuition only works for the nonsocial environment. For the

social environment, there is instead a strong plausibility that brains co-evolved, and so cognition indeed could evolve through changes across multiple brains.

To summarize: my current concept of "cognition", however ill-defined, is squarely centered on the brain. But I have not made a serious attempt to revise this, and it is possible that such a revision would result in a concept with more utility. The argument that causal density specifies nature's joints for a cognitive system is problematic because actual physical density of connections is probably not the right metric. These considerations lead to the conclusions of the next section, on which all three of us agree.

6. Suggestions for social neuroscience

In writing this article, all three of us acknowledge that understanding the brain and cognition is incredibly hard. All approaches should have the provisional status of "hypothesis" – something the IBH explicitly does have, but standard information processing views usually do not. We agree that historical views of cognition as computation over representations are unlikely to adequately describe how brains work. Rather than defining cognition as that kind of information processing that is unique to brains, we would prefer to think of cognition as involving brains in some way yet to be fully understood, possibly including causal relations between brains, bodies, other people, and even the nonsocial environment. Evolution has made use of whatever substrates are available to generate flexible behavior, and we simply do not know yet what those substrates are.

The IBH can be seen as making a practical claim for scientists, namely that there is a more compact explanation of human social behavior if we adopt the interactive stance than if we stick with the classical input-output stance. Consider how far one could push the classical "brain-in-a-vat" thought experiment. There have been recent experiments using optogenetics in mice that manipulate brain activity so precisely that they literally reconstitute the pattern of neuronal activity that would have been evoked by encoding an actual sensory stimulus [51]. Such experimentally created patterns of activity in the brain cause the mouse to behave as if it remembered an actual stimulus. While this experiment seems to show that we can understand cognition and behavior as divorced from the environment, it actually points to the value of the IBH as a framework to understand what is happening. Suppose the experiment attempted to re-create the pattern of activity involved in an actual, reciprocal, social encounter. It quickly becomes apparent that to do so would require mimicking the other animal as a social stimulus. But since the other animal responds to our experimental mouse, this is not a fixed input, but rather a complex, time-varying input embedded in causal loops with the very behavior we wish to experimentally control. Our surrogate "input pattern" ends up being not only extremely complex, but in fact cannot be specified in the absence of an analysis of the first mouse as involved in a socially coupled interaction, which itself can show emergent dynamical patterns that do not reduce to the activity of the mice. This conclusion is consonant with Cosmelli and Thompson's suggestions about the brain-in-a-vat thought experiment [18]. The upshot is that we may be able to describe some aspects of cognition reasonably well as input-output transformations by the brain, whereas others cannot be so described. Social cognition typically may be of the latter kind.

This leads us to consider the practical issue of which are the criteria for delineating the systems under study. One possible reading of the IBH could be that everything matters, and so there is no right decomposition into causal systems that could illuminate a particular instance of social cognition. This is certainly not the intended reading.[2] To highlight the embeddedness of brain activity within a behaving and interacting body is not to render social neuroscience a hopeless endeavour. It is to raise awareness that certain assumptions, such as the assumption of decomposability, can be problematic if formulated uncritically. There exist many experimental approaches that would allow the simultaneous study and manipulation of neural and interactive dynamics, as we have mentioned.

---

[2] The situation is not unlike other debates in biology. Arguments for the importance of non-genetic factors in evolution and development (e.g [52,53]) are often met with the criticism that one cannot study every conceivable causal factor scientifically. But there are positive counter-proposals that distinguish between different causal roles. For instance, Woodward [54] suggests that one can discriminate between causal factors according to different criteria such as their stability or non-contingency, specificity, and appropriateness for the level of explanation. Manipulability could be another factor for this kind of consideration.

Besides these practical claims, the IBH can also be seen as a foundational claim about what grounds social cognition. From input-output transformations alone meaning cannot emerge; for what do the inputs and outputs stand for? The IBH assigns meaning from the wider perspective of 1) social interaction, and 2) the wider enactive theory about cognition as sense-making that it forms part of (e.g. [55]). From an enactive perspective, meaning emerges in virtue of historical and concurrent patterns of interaction. That something like this must happen, for example, as an infant learns her very first words seems uncontentious. Associating the word "book" with seeing a book presumably works by the infant and another person both looking at the book and saying and hearing the word "book". But why does she learn the word "book"? Why is it meaningful to her at all? Here, the wider framework of participatory sense-making out of which the IBH grew makes claims about how cognizers encounter the world as meaningful (which is how enactivists define cognition) (43; 55). Certainly, there is room for rich further debate here. The key future challenge for the enactive approach is to develop further concepts and hypotheses, and to continue to articulate them in ways that make contact with other frameworks. In this way, concepts like "participatory sense-making" can be articulated into domain-specific claims, hypotheses, and explanations that relate to conceptions of meaning, as they vary between relevant disciplines. Formulating the IBH is an attempt to do precisely this for the field of social neuroscience.

A closing question is where to find a home for social neuroscience in all of this. The difficulty arises when social neuroscience attempts to study that which *underlies* behaviour and cognition, when it attempts to explain how meaning is generated, and why social cognition and social behavior exhibit particular forms and features. After all, social cognition shows substantial differences if we compare a dog, a chimpanzee, or for that matter people from different cultures or at different ages. How can we explain these differences – differences that render social behavior meaningful for individuals of each species, culture, and epoch, but less meaningful as we cross between these. Perhaps the largest contribution of the IBH to social neuroscience is to show that it is impossible to answer these questions if the only data we entertain are from a single adult brain in one species. That is, we need to consider species-typical social interactions not just in the context of meaningful situations, but also in the context of evolution and development. Echoing the well-known ethological refrain that "nothing makes sense in biology except in the light of evolution" [56, p.125], we would urge that social neuroscience should incorporate at least comparative neuroscience, developmental neuroscience, and input from sciences that study social interaction into its domain of study.

Additional Information

Acknowledgments
We thank Guillaume Dumas, Frederick Eberhardt, Riitta Hari, and the reviewers for their comments on the manuscript.

Authors' Contributions
EDP and HDJ are the primary authors of sections 2 and 4; RA is the primary author of sections 3 and 5; all authors contributed equally to sections 1 and 6, and all authors provided substantial input to all sections of the paper.

Competing Interests
We have no competing interests.

Funding
HDJ is funded by a Ramón y Cajal Fellowship, RYC-2013-14583. RA was supported in part by a Conte Center grant from the National Institute of Mental Health (USA).

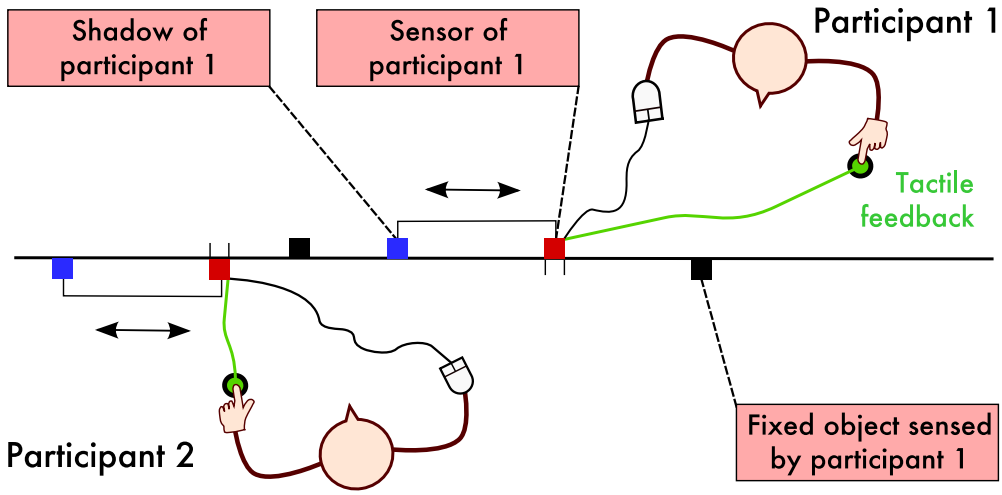
References

1. Di Paolo, E.A. & De Jaegher, H. 2012 The Interactive Brain Hypothesis. *Frontiers in Human Neuroscience* **6**. (doi:10.3389/fnhum.2012.00163).
2. De Jaegher, H., Di Paolo, E.A. & Gallagher, S. 2010 Can social interaction constitute social cognition? *Trends in Cognitive Sciences* **14**, 441-447. (doi:doi:10.1016/j.tics.2010.06.009).
3. Harlow, H.F. & Harlow, M.K. 1962 Social deprivation in monkeys. *Scientific American* **207**, 136-146.
4. Auvray, M., Lenay, C. & Stewart, J. 2009 Perceptual interactions in a minimalist virtual environment. *New Ideas in Psychology* **27**, 32-47.
5. Clark, A. 1997 *Being There: Putting Brain, Body and World Together Again*. Cambridge, MA, MIT Press.
6. Kelso, J.A.S. 1995 *Dynamic Patterns: The Self-Organization of Brain and Behaviour*. Cambridge, MA, MIT Press.
7. Noë, A. 2004 *Action in Perception*. Cambridge, MA, MIT Press.
8. Findlay, J.M. & Gilchrist, I.D. 2003 *Active vision: the psychology of looking and seeing*. Oxford, Oxford University Press.
9. Przyrembel, M., Smallwood, J., Pauen, M. & Singer, T. 2012 Illuminating the dark matter of social neuroscience. *Frontiers in Human Neuroscience* **6**, 190.
10. Koster-Hale, J. & Saxe, R. 2013 Theory of Mind: a neural prediction problem. *Neuron* **79**, 836.
11. Clark, A. 2013 Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences* **36**, 181-204.
12. Noë, A. 2010 *Out of our heads: why you are not your brain, and other lessons from the biology of consciousness*. London, Macmillan Press.
13. Clark, A. & Chalmers, D.J. 1998 The extended mind. *Analysis* **58**, 7-19.
14. [1] Rowlands, M. 2003 *Externalism: putting mind and world back together again*, Acumen/McGill-Queens University Press.
15. Byrge, L., Sporns, O. & Smith, L.B. 2014 Developmental process emerges from extended brain–body–behavior networks. *Trends in Cognitive Sciences* **18**, 395-403.
16. Simon, H.A. 1962 The architecture of complexity. *Proceedings of the American Philosophical Society* **106**, 467-482.
17. Anderson, M.L., Richardson, M.J. & Chemero, A. 2012 Eroding the boundaries of cognition: Implications of embodiment. *Topics in Cognitive Science* **4**, 717-730.
18. Cosmelli, D. & Thompson, E. 2011 Brain in a vat or body in a world: Brainbound versus enactive views of experience. *Philosophical Topics* **39**, 163-180.
19. Babiloni, F. & Astolfi, L. 2014 Social neuroscience and hyperscanning techniques: Past, present and future. *Neuroscience and Biobehavioral Reviews* **44**, 76–93.
20. Simon, H.A. 1969/1996 *The Sciences of the Artificial*. Cambridge, Mass., MIT Press.
21. Dumas, G., Nadel, J., Soussignan, R., Martinerie, J. & Garnero, L. 2010 Inter-brain synchronization during social interaction. *PLoS ONE* **5**, e12166.
22. Canolty, R.T., Edwards, E. & Dalal, S.S., Soltani, M., Nagarajan, S. S., Kirsch, H. E., et al. 2006 High gamma power is phase-locked to theta oscillations in human neocortex. *Science* **313**, 1626–1629.
23. Womelsdorf, T., Schoeffelen, J.-M., Oostenveld, R., Singer, W., Desimone, R., Engel, A. K., et al. . 2007 Modulation of neuronal interactions through neuronal synchronization. . *Science* **316**, 1609–1612.
24. Buzsáki, G. & Draguhn, A. 2004 Neuronal oscillations in cortical networks. *Science* **304**, 1926-1929.
25. Le van Quyen, M. 2011 The brainweb of cross-scale interactions. *New Ideas in Psychology* **9**, 57–63.
26. He, B.J., Zempel, J.M., Snyder, A.Z. & Raichle, M.E. 2010 The temporal structures and functional significance of scale-free brain activity. *Neuron* **66**, 353-369.
27 Buschman, T.J., Miller, E. K. 2007 Top-down versus bottom-up control of attention in the prefrontal and posterior parietal cortices. *Science* **315**, 1860-1862.
28. Saalmann YB, P.I., Vidyasagar TR 2007 Neural mechanisms of visual attention: how top-down feedback highlights relevant locations. *Science* **316**, 1612-1615.
29. Lakatos, P., Karmos, G., Mehta, A. D., Ulbert, I., Schroeder, C. E. 2008 Entrainment of neuronal oscillations as a mechanism of attentional selection. *Science* **320**, 110–113.
30. Astolfi, L., Toppi, J., de Vico Fallani, F., Vecchiato, G., Salinari, S., Mattia, D., Cincotti, F., Babiloni, F. 2010 Neuroelectrical hyperscanning measures simultaneous brain activity in humans. *Brain Topography* **23**, 243–256.
31. Astolfi, L., Toppi, J., de Vico Fallani, F., Vecchiato, G., Cincotti, F., Wilke, C. T., Yuan, H., Mattia, D., Salinari, S., He, B., Babiloni, F. 2011 Imaging the social brain by simultaneous hyperscanning during subject interaction. *IEEE Intelligent Systems* **26**, 38-45.
32. Dumas, G., Kelso, J.A.S. & Nadel, J. 2014 Tackling the social cognition paradox through multi-scale approaches. *Frontiers in Psychology* **5**. (doi:10.3389/fpsyg.2014.00882).
33. Hari, R. & Kujala, M.V. 2009 Brain basis of human social interaction: From concepts to brain imaging. *Physiological Reviews* **89**, 453-479.
34. Kelso, J.A.S., Dumas, G., Tognoli, E. 2013 Outline of a general theory of behavior and brain coordination. *Neural Networks* **37**, 120–131. (doi:10.1016/j.neunet.2012.09.003).
35. Van Orden, G.C., Holden, J.G. & Turvey, M.T. 2003 Self-Organization of Cognitive Performance. *Journal of Experimental Psychology* **132**, 331-350.
36. Dotov, D.G., Nie, L., Chemero, A. 2010 A demonstration of the transition from ready-to-hand to unready-to-hand. *PLoS ONE* **5**, e9433.
37. Shockley, K., Santana, M.-V. & Fowler, C.A. 2003 Mutual interpersonal postural constraints are involved in cooperative conversation. *Journal of Experimental Psychology: Human Perception and Performance* **29**, 326-332.
38. Richardson, M.J., Marsh, K.L., Schmidt, R.C. 2010 Challenging egocentric notions of perceiving, acting, and knowing. (New York, Guildford.
39. Riley, M.A., Richardson, M., Shockley, K. & Ramenzoni, V.C. 2011 Interpersonal synergies. *Frontiers in Psychology* **2**. (doi:10.3389/fpsyg.2011.00038).
40. [1] Bedia, M.G., Aguilera, M., Gomez, T., Larrode, D.G. & Seron, F. 2014 Quantifying long-range correlations and 1/f patterns in a minimal experiment of social interaction. *Frontiers in Psychology* **5**. (doi:10.3389/fpsyg.2014.01281).
41. Clark, A. Thornton, C. (1997). Trading spaces: Computation, representation, and the limits of uninformed learning, *Behavioral and Brain Sciences*, 20 (1997) 57–66.
42. Froese, T., Iizuka, H. & Ikegami, T. 2014 Embodied social interaction constitutes social cognition in pairs of humans: A minimalist virtual reality experiment. *Scientific Reports* **4**. (doi:http://dx.doi.org/10.1038/srep03672).
43. De Jaegher, H. & Di Paolo, E. 2007 Participatory Sense-Making: An enactive approach to social cognition. *Phenomenology and the Cognitive Sciences* **6**, 485-507. (doi:DOI 10.1007/s11097-007-9076-9).
44. Granic, I. 2000 The self-organization of parent-child relations: beyond bidirectional models. In *Emotion, Development, and Self-Organization. Dynamic Systems Approaches to Emotional Development* (eds. M.D. Lewis & I. Granic), pp. 267-297. Cambridge, Cambridge University Press.
45. Shergill, S.S., Bays, P. M., Frith, C. D., Wolpert, D. M. 2003 Two eyes for an eye: The neuroscience of force escalation. *Science* **301**, 187.
46. Bateson, G. 1936 *Naven*. Stanford, CA, Stanford University Press.
47. Watzlawick, P., Helmick Beavin, J. & Jackson, D.D. 1967 *Pragmatics of Human Communication. A Study of Interactional Patterns, Pathologies, and Paradoxes*. New York/London, Norton; 294 p.
48. Stern, D.N. 1977/2002 *The First Relationship: Infant and Mother*. 2nd ed. London, Harvard University Press.
49. Fogel, A. 1993 *Developing through relationships: Origins of communication, self and culture*. London, Harvester Wheatsheaf.
50. Searle, J. 1990 Consciousness, explanatory inversion, and cognitive science. *Behavioral and Brain Sciences* 13, 585-642.
51. Ramirez, S., Liu, X., Lin P.A., et al. 2013 Creating a false memory in the hippocampus. *Science* **341**, 387-391.
52. Oyama, S. 1985 *The Ontogeny of Information: Developmental Systems and Evolution*. Cambridge, Cambridge University Press.
53. Jablonka, E. & Lamb, M.J. 2005 Evolution in Four Dimensions: Genetic, Epigenetic, Behavioral, and Symbolic Variation in the History of Life. (Cambridge, MA.
54. Woodward, J. 2010 Causation in biology: stability, specificity, and the choice of levels of explanation. *Biology & Philosophy* **25**, 287-318.
55. Thompson, E. 2007 *Mind in Life: Biology, Phenomenology, and the Sciences of Mind*. Cambridge, MA, Harvard University Press.
56. Dobzhansky, T. (1973). Nothing in Biology makes sense except in the light of evolution. *American Biology Teacher* 35: 125-129.

Figure 1 caption

Figure 1: Set-up for perceptual crossing experiment. Both participants are isolated, each controlling the position of a sensor along a shared virtual 1-D line using a computer mouse. The squares on each side of the line represent the objects that can be sensed by each participant respectively. Objects are identical in size. When the sensor touches an object the participant gets a tactile feedback on the finger (green circle). Each participant can sense only three objects, a static one (black square), the sensor of the other participant (red square) and a 'shadow' object that copies exactly the movement of the other's sensor at a fixed distance (blue square).

Figure 1


Shadow of participant 1
Sensor of participant 1
Participant 1
Tactile feedback
Participant 2
Fixed object sensed by participant 1