SRI International

CSL Technical Report • July 1, 2023

On Computational Mechanisms for Shared Intentionality And Speculation on Rationality and Consciousness

John Rushby
Computer Science Laboratory
SRI International, Menlo Park CA 94025 USA

This research was sponsored by SRI International and by my retirement plan

Computer Science Laboratory • 333 Ravenswood Ave. • Menlo Park, CA 94025 • (650) 859-2000 • Facsimile: that's last century

Contents

List of Figures

On Computational Mechanisms for Shared Intentionality And Speculation on Rationality and Consciousness

John Rushby
Computer Science Laboratory
SRI International, Menlo Park CA 94025 USA
Rushby@csl.sri.com

Abstract

A singular attribute of humankind is our ability to undertake novel, cooperative behavior, or teamwork. This requires that we can communicate goals, plans, and ideas between the brains of individuals to create *shared intentionality*. I adopt the view that the brain performs computation, then, using the information processing model of David Marr, I derive necessary characteristics of basic mechanisms to enable shared intentionality between prelinguistic computational agents.

More speculatively, I suggest the mechanisms derived by this thought experiment apply to humans and extend to provide explanations for human rationality and aspects of intentional and phenomenal consciousness that accord with observation. This yields what I call the Shared Intentionality First Theory (SIFT) for language, rationality and consciousness.

The significance of shared intentionality has been recognized and advocated previously, but typically from a sociological or behavioral point of view. SIFT complements prior work by applying a computer science perspective to the underlying mechanisms.

1 Introduction

About 20 years ago, our neighboring AI laboratory engaged in a project named *Centibots* [86]. The Centibots were small mobile robots, each about a foot cube, and there were a hundred of them, hence the name (see Figure 1). They were equipped with cameras and other sensors, and had a limited ability to communicate with each other.

The idea was that Centibots would be deposited in a building and would then spread out and collaboratively develop a map of the internal layout. To do this, each Centibot would exchange some state information with others in its vicinity and apply this, together with information from its own sensors, to develop and execute a continually updated plan to achieve its part of the overall goal. Observe that the function and behavior of the Centibots was not unlike that of social animals, such as bees and wolves.

Figure 1: Centibots

Let us now leave the historical Centibots, and conduct a thought experiment. Suppose we took one of the Centibots and reprogrammed it with some additional planning capabilities and a new goal: to form, together with its cohorts, a line that can guide humans to an exit in case of emergency. It is obvious that no progress can be made in this endeavor unless there is some way to share the new goal with the other Centibots. But the standard Centibots do not have this capability: their communications "language" is limited to the information needed for their original task, rather like the "dances" of bees and the howls and body postures of wolves. Furthermore, sharing the goal might not be enough: the standard Centibot planner might not be able to come up with local actions to achieve the new goal, so it may be necessary for the modified Centibot also to communicate some "hints" or "ideas" to augment the local planners. This cannot be accomplished by simply watching the modified Centibot executing part of the goal (e.g., standing by an exit).

We see that to get from preprogrammed collaborative behavior to the ability to jointly undertake *new* tasks requires additional capabilities that allow novel information to be communicated from one Centibot to another in such a way that the recipient comes to share some of the goals and planning elements of the sender.

Just as the basic Centibots can serve as crude models for social insects and animals, so the hypothetically augmented Centibots can serve as models for animals that can engage in novel forms of collaborative behavior: that is, in teamwork. The additional capabilities of the augmented Centibots are an abstracted characterization of what, in animals, is termed "shared intentionality": that is, the ability to communicate and share similar mental states that can drive similar behavior [139].[1]

I believe that modern humans are the only animals that engage in full-fledged teamwork[2] and this (and its larger manifestation as culture [68,130]) is the reason we dominate the world. Furthermore, I propose that our other defining capabilities—language, rationality, consciousness—arose from the mechanisms that enable teamwork: that is, from shared intentionality. I will develop a computer science description of how shared intentionality might be constructed in hypothesized post-Centibot computational agents. I will then argue that these mechanisms extend to real humans[3] and provide, almost immediately, the capabilities for rational deliberation; I speculate that awareness of these processes constitutes intentional consciousness. Phenomenal consciousness arises because we then need the ability to communicate the content of our sense experience.

I refer to this collection of computer science, deduction, and speculation as the "Shared Intentionality First" Theory of rationality and consciousness (SIFT) and I argue that it explains otherwise puzzling aspects of these faculties and is evolutionarily plausible. SIFT is more abstract than other theories of consciousness: it concerns the purpose, strategy, and architecture of mental mechanisms, not their biological implementation, and so it is compatible with, or provides context for, several other theories of brain organization and consciousness, including predictive processing [28,69], higher-order thought [55,124] and its related notion of metacognition [17], dual-process theories [48,82], and global workspace theories [5,32]. SIFT

[1] Intentionality is the property of computational or mental states being *about* or *directed toward* something but we (or computational agents) can have many different mental attitudes toward that something besides intentions: we may, for example, believe it, fear it, prefer it to something else, or want it, and so on. The philosopher's jargon "intentional" (due to Franz Brentano) comes by way of translation from German and should not be construed to refer specifically to "intentions."

[2] For example, "it is inconceivable that you would ever see two chimpanzees carrying a log together" [66, quoting Tomasello on page 238].

[3] I take it as a given that the function of the brain is to perform computations. This is not a metaphor or simile, as when in earlier times the brain was said to be "like" a clock; we are saying it *is* a computer. This confuses some who are familiar only with desktop computers: the brain is a computer in the sense that it performs computations, but it is obviously organized, implemented, and deployed completely differently than a desktop computer; for a mechanistic analogy, think of the computational system of a self-driving car and its integration with perception and actuation.

also brings a different perspective to prior work on the rôle of shared intentionality in the development of human cognition and language [9, 10, 136, 137]: in particular, SIFT's computer science focus on the underlying mechanisms complements prior work on the behavioral and psychological attributes of shared intentionality and suggests how these form an integrated cognitive "package."

The paper is organized as follows: the next section considers the architecture of mechanisms for construction of shared intentionality in humanoid computational agents. Although this description is driven by intuitions, mechanisms, and technologies from computer science and computational models of perception, it is written at a tutorial level and is intended to be accessible to all. I then propose that real humans have similar mechanisms; this is followed by sections on rationality, intentional consciousness, and phenomenal consciousness, respectively. I then briefly consider the archaeological record and the evolutionary plausibility of SIFT, followed by comparison with some other theories and with attempts to create artificial consciousness and conclude with a summary of fundamental claims and acknowledgments.

2 The Construction of Shared Intentionality

I will couch my construction of shared intentionality in terms of hypothesized human-like computational agents (i.e., mobile robots) that operate in the open world and could be considered humanoid "descendants" of Centibots: by "humanoid" I mean their sensors and actuators resemble human senses and limbs, their sense interpretation, planning, and execution capabilities are powerful and comparable to those of the human subconscious, and they are autonomous (i.e., set their own goals),[4] but they lack mechanisms for novel collaborative behavior: no language or other means to exchange state information or ideas beyond those preprogrammed for specific tasks as in the original Centibots. I make these choices because I will later argue that these agents, when equipped with shared intentionality, serve as models for real humans, and for this reason I will sometimes use anthropomorphic terms in my discussion.

In particular, let us imagine an individual agent of this type faced with a problem: crossing a small ravine. Thanks to its large modeling, search, and planning capabilities our individual—let's call it/her Alice—might invent a novel behavior: use a fallen tree trunk or log to create a bridge across the ravine. Now let us suppose the log is too large and heavy for Alice to move into place. A second agent—let's

[4] I recognize the difficulties here: what exactly are the capabilities of the human subconscious, and how can we have autonomy without encountering the philosophical problems of free will? However, for the purposes required here, I believe these topics can be set aside (i.e., we can suspend disbelief) and that some approximate interpretation is sufficient.

call it/him Bob[5]—may watch her struggling with the log but, having no prior experience of log-bridges, he will not understand what is going on any more than would a dog or a human infant, and will render no useful assistance (see Figure 2).[6]

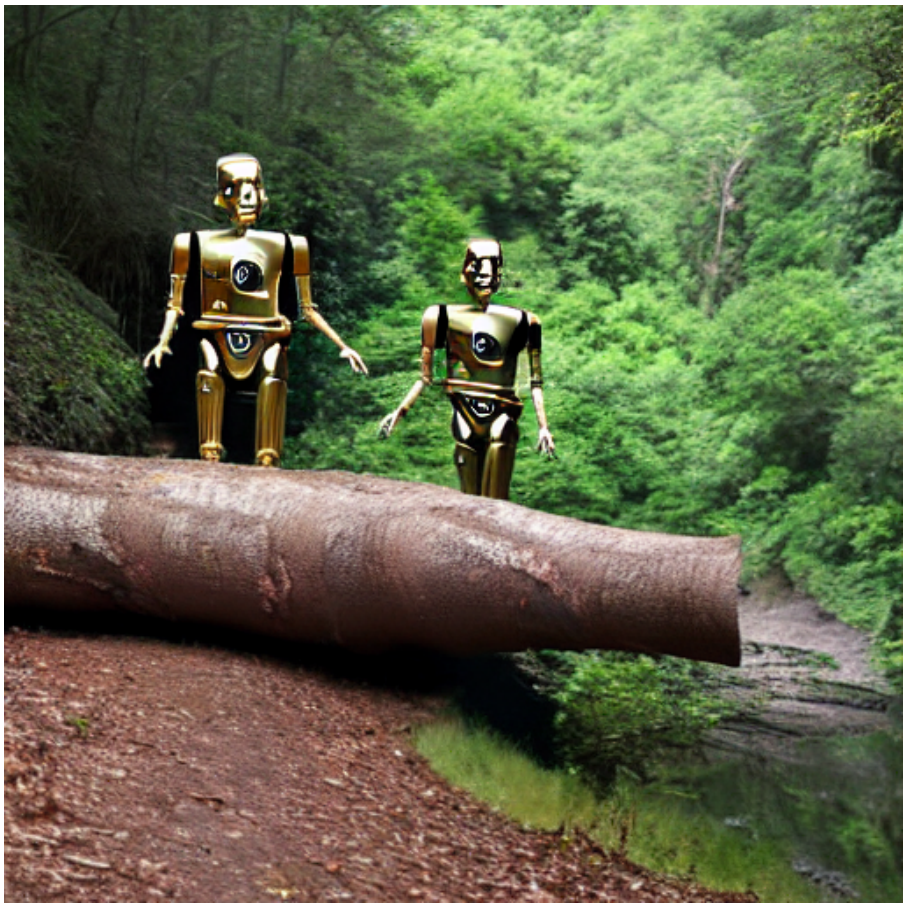

Image created by DreamStudio, CC0 1.0 Universal Public Domain Dedication

Figure 2: Two Humanoid Agents and a Log-Bridge

Notice that the behavior required in cooperative construction of a log-bridge is quite different than simply copying from observation. If Alice has a technique for using stones to crack nuts, Bob may be able to learn this by simply observing and

[5]Alice and Bob are a standard trope for describing distributed algorithms in computer science (see Wikipedia [146]). Their appearance here does not indicate any specific debt to other papers that happen to have adopted the same usage.

[6]It is possible that Bob will copy Alice's activity as a preprogrammed form of cooperation. However, not understanding the purpose, he might very likely push the log in the wrong direction, or push a different log.

copying her behavior. Alice does not need to explicitly communicate the skill to Bob (indeed, she need not be aware that she has the skill, as seems to be the case with chimpanzees), and Bob does not need to infer the goal and technique: they are there before him. But the log-bridge exists only in Alice's head: Bob must somehow infer this goal and Alice must help him do so, and likewise the plan to achieve it by moving a specific log into place. To get truly cooperative behavior on a novel task, the individuals must have "shared intentionality": that is, similar computational or mental states that can drive similar behavior [139].

In order to examine mechanisms that could bring this about, I adopt and adapt the three-stage "information processing" model of David Marr [97]. The model concerns human (or animal) cognition but presupposes that mental operations are information processing tasks (i.e., computations), so it applies directly to our hypothesized computational agents. The first step or stage in understanding such a computation is to deduce its goal and the strategy by which it can be carried out. The second stage considers the representations that can be employed for the inputs and outputs of the computation and the mechanism or "algorithm" that can transform the one into the other. The third and final stage considers the implementation that can physically realize the representation and algorithm: digital circuits and software in the case of our agents, and neurons and other biology in the case of humans and animals.

The goal of the computation under consideration is the creation of shared intentionality and I already staked out a position when I suggested this requires the parties to achieve "similar computational or mental states that can drive similar behavior." Thus, I will suppose that the immediate goal is to recreate in Bob some aspects of Alice's computational state (specifically, that associated with the "idea" of a log-bridge), so that his planner has access to similar information and may use this to generate usefully cooperative behavior [15]. At this point, we must adjust Marr's model a little because the computation underlying shared intentionality must surely be a *distributed* one: some of it will be performed by Alice and some by Bob, and something will be transferred between them. We now need to consider what are the separate computations, and what is transferred.

Alice and Bob might be constructed differently (e.g., Alice might be programmed in Java and Bob in Python) so we cannot simply copy some bag of bytes from Alice's state to Bob's and expect it to have a useful effect. Similarly, if Alice and Bob were biological entities, their brains, even as conspecifics, will have grown and developed somewhat differently due to their individual genetics, physiology, and experiences, and so the transfer mechanism cannot directly reconstruct part of Alice's neural state (i.e., the electrical and chemical activity in some specific cluster of neurons) in Bob's brain since they will be "wired up" differently at the neural level.

Thus, what is transferred cannot be a representation or description of the implementation level of the agent's computational state. Instead, it must be abstracted

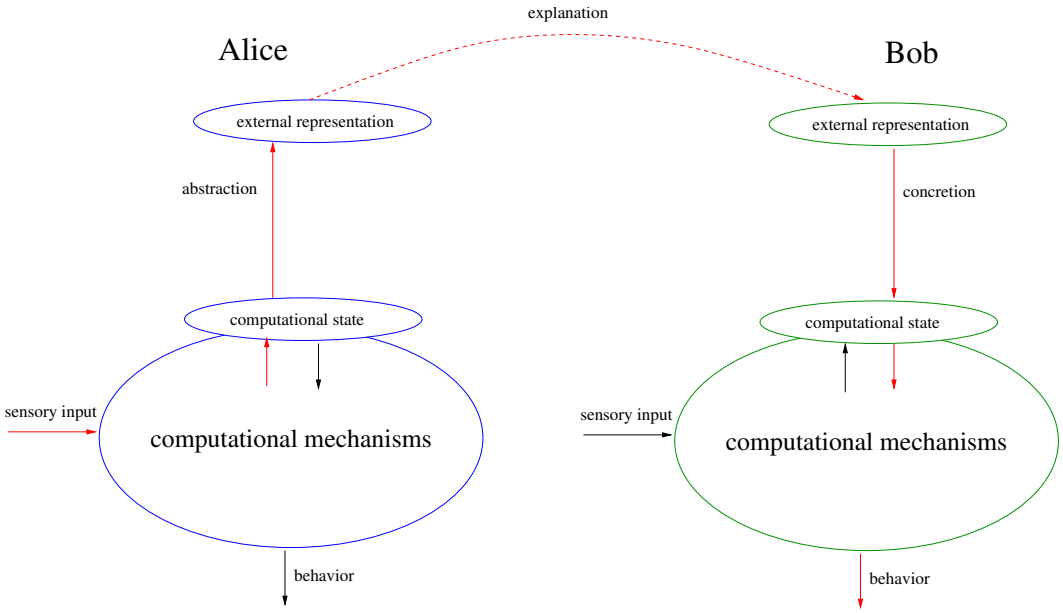

Figure 3: Architecture of Mechanism Necessary to Create Shared Intentionality

into some representation that is common to both Alice and Bob. It needs to be abstracted for two reasons: first, it must be feasible to communicate it (e.g., by demonstration, mime, or—later—language), so it must be succinct; second, it must be common to all participants. In outline, the strategy for shared intentionality will then be as follows: Alice computes an abstraction of relevant aspects of her low-level computational state into a form I call the "external representation"; this is communicated to Bob (I consider how this is done below), whose computation inverts the abstraction that created this representation, an operation referred to as "concretion," and thereby enriches his computational state so that it now contains information and "ideas" similar to Alice's and this may lead him to perform usefully cooperative behavior. Figure 3 portrays this design, and the flow of information (red arrows)[7] from Alice (on the left, in blue) to Bob (on the right, in green). Observe that concretion is performed by Bob, and is therefore able to target the low-level representation used for his computational state.

We have seen that the external representation cannot be couched in terms of the physical computational or mental state—that is, in terms of bits and bytes, or neurons, connections, chemistry, and signals. Instead, it must employ some more

[7]The external flow labeled "explanation" is shown dotted because this is a virtual, rather than physical, flow of information; the physical flow is accomplished by Alice generating behavior that is sensed by Bob.

abstract vocabulary[8] that is common to Alice and Bob. If the agents were built on current technology, Alice and Bob's developers would likely use unsupervised machine learning to provide them with means to interpret their perceptions and could choose methods that will "chunk" their world similarly (e.g., [135]), even though they might be developed separately. Anthropomorphically, I will refer to these chunks as "concepts."

The external representation might be no more than a string of concepts, but it will be more effective if these can be linked together in a way that indicates sequencing, intent, or causation. Thus, in addition to sensed objects (i.e., prelinguistic nouns), chunking should recognize and the vocabulary should include actions (verbs) and spatial and temporal relations (prepositions). The abstraction mechanism then constructs a "story" or, as I will say, an *explanation* in this vocabulary that suggests a state of the world and actions on it whose concretion matches the salient part of Alice's mental state (that is, her "idea"). Here, it is likely that Bob is missing the concept "bridge," so Alice might employ the concepts "walk," "on top of," "fallen tree," "across," and "ravine" and her explanation will link these together in sequence, and perhaps indicate intent: "in order to" "go to" "other side."[9]

The means whereby Alice conveys the explanation to Bob might be demonstration, mime, or signs and sounds (repurposed from the communications built-in for preprogrammed interactions) that by convention are associated with specific concepts. For example, she might use a twig to scrape a small ditch in the dirt, then lay the twig across it and "walk" her fingers across, and then point to the ravine and indicate the selected log. Bob will watch this mime and his sensory-processing faculties must recognize that it has symbolic or conceptual content, extract the concepts and explanation, and then concretize them so that they are available to his computational state and mechanisms.

We have now applied the first two stages of Marr's model: we have postulated the purpose and strategy of the computations performed by Alice and Bob in achieving shared intentionality (i.e., use of abstraction/concretion), and of the representations employed and communicated between them (i.e., explanations over concepts). Now let us consider how Alice might perform abstraction and thereby derive something of the third or algorithmic stage of description. Some of the mechanisms I propose, and some that I suppose are already present in our agents, may seem rather arbitrary, but there is a purpose behind these choices: they are based on those known or hypothesized to operate in the human brain.

[8] Computer scientists would generally use the term "ontology" here, but that usage is nonstandard and somewhat idiosyncratic to the field, so I prefer the more neutral term "vocabulary."

[9] Those aware of dialog systems based on "Large Language Models," such as ChatGPT and its cohorts [112], may wonder why I do not invoke a reduced form of this technology here. The reason is that our hypothesized agents are prelinguistic: we are attempting to discover where language comes from.

The goal for our algorithm is to construct an explanation—a succinct external communication—whose concretion by Bob will reproduce parts of Alice's computational state. If Alice is to construct such an explanation, she surely needs to do it using an estimate of Bob's concretion operation: hence, she must have what philosophers and psychologists call a "theory of mind" [98, 116].[10] For humans, researchers divide "social information processing" into processes that are relatively automatic and driven by stimuli, versus those that are more deliberative and controlled. These distinctions are reflected in the neural structures that underlie social cognition [1]. We suppose that processes similar to the automatic ones are part of Alice's basic computations and lead her to suppose that Bob is similar to her (so it is worth trying to communicate with him), and the "more deliberative and controlled" processes are part of the mechanism whose structure we are attempting to deduce.

It is generally understood that any mechanical or living entity that interacts effectively with some aspect of the world (its *environment*) must have a model of that environment [29, 47].[11] For example, the construction and maintenance of an "adequately correct" model of its environment is currently the central problem in design and assurance of autonomous systems such as self-driving cars [80]. Bob is part of Alice's environment, so we may suppose that she has a model of Bob's state of knowledge and beliefs[12] and will use this to guide construction of her explanation: she will use a different explanation if she believes Bob already has the concept of log-bridges than if he does not. So now we ask: how does Alice use her model of Bob's computational state to guide her abstraction?

Let s denote Alice's computational state; her abstraction operation $Alice_A$ needs to take the relevant part of her state (i.e., that concerning her idea for a log-bridge), which we will denote $idea(s)$, and deliver candidate explanations. We suppose that the "more deliberative and controlled" aspects of Alice's theory of mind for Bob reside with the new computational mechanism that we are attempting to construct, so Alice's abstraction operation needs the potential to offer many explanations, so that the new mechanism can pick the one, e, that will be most effective. Thus we have

$$e \in Alice_A(idea(s)) \tag{1}$$

[10] Theory of mind, also known as "mindreading" [6], should not be confused with shared intentionality. Mindreading infers another agent's beliefs and intentions by observation of its behavior in its environment (e.g., by simulating its point of view, or by applying deduction [22]); shared intentionality involves transfer of internal "ideas" that cannot be directly observed or inferred: their communication requires deliberate, symbolic actions, such as Alice's mime.

[11] Conant and Ashby explicitly recognized this must apply to the brain, which seems remarkably prescient for 1970: "The theorem has the interesting corollary that the living brain, so far as it is to be successful and efficient as a regulator for survival, must proceed, in learning, by the formation of a model (or models) of its environment" [29].

[12] Beliefs, Desires, and Intentions (BDI) are a standard way of organizing some aspects of an AI agent [118]. Knowledge is understood as true belief.

and we ask how $Alice_A$ is constructed and how a suitable e is selected.

Now, Alice has her own concretion faculty (for use when she is the receiver) and I temporarily propose[13] that this can be parameterized by models for different "points of view" (i.e., theories of mind) and thereby simulate (approximately) Bob's concretion. For simplicity of exposition, assume that Bob's concretion operation and Alice's simulation of it are deterministic functions. We denote Alice's operation by $Alice_C(Bob, e)$, where $Alice_C$ is Alice's concretion function, the argument Bob indicates this application is parameterized by her model of Bob, and e is an explanation. This function delivers a concretized version of e in the computational form employed by Alice that represents her estimate of Bob's interpretation of e. We can use the keyword *myself* to indicate application of the native (unparameterized) concretion function, so that Bob's native concretion function is $Bob_C(myself, e)$, and this delivers the computational representation of e used by Bob.

What we want is that $Bob_C(myself, e)$ is an augmentation to Bob's computational state that is similar in effect to the relevant part of Alice's state $idea(s)$. We cannot require $Bob_C(myself, e) \approx idea(s)$ because the left side uses Bob's computational representation, while the right side uses Alice's. But what we can do is require that Alice's simulation of Bob's concretion delivers a value (which will be in her representation) that is close to her state $idea(s)$. That is

$$Alice_C(Bob, e) \approx idea(s). \tag{2}$$

If we use a function *error* to measure divergence between the left and right sides of (2) then we want to choose an e that minimizes this divergence. That is, combining (1) and (2):

> Alice's explanation for Bob of her idea $idea(s)$ is e that minimizes
>
> $$error(Alice_C(Bob, e), idea(s)) \tag{3}$$
>
> over all $e \in Alice_A(idea(s))$.

This constraint ensures that Alice chooses a good explanation, given her model of Bob. We can suppose that pre-existing mechanisms allowed her to build a model for Bob (see, e.g., [142]) and, if necessary, to refine it in a failure-driven "dialog" should it prove inadequate.

Solving constraint satisfaction problems such as (3) usually requires some form of optimizing search. Fortunately, we may suppose that Alice already has the mechanisms for such search. As we noted earlier, any autonomous agent must use information from its sensors to build and refine a model of its environment. The untutored view is that the model is built from the sensors "bottom up," as when

[13] I say "temporarily" because I will later suggest a more realistic mechanism, but we do not yet have the context for its introduction.

machine learning is used to build the "detected objects list" from the cameras in a self-driving car. However, this approach is prone to error and instability, not least because it operates *anti-causally* [83]. A better alternative turns things around and uses the model to generate or *predict* observations and then uses the resulting *prediction error* to refine the model [80].[14] This can be mechanized using techniques known as Variational Bayes [46] in which inference is achieved via optimization: here, the model is probabilistic (e.g., probability distributions over attributes of items in the detected objects list), predictions correspond to Bayesian *priors*, prediction errors encode observations, and the Variational Bayes algorithm constructs an update to the model (the Bayesian *posterior*) that approximates that needed to minimize future prediction error.[15] Similar processes are thought to operate (across a hierarchy of models) in the brain (we discuss this later) where they are described as *Predictive Processing* (PP) [145] and I will use the same term here.

We suppose that our agents' basic computational mechanisms (in particular, their perception systems) use PP to build models of the environment and these are represented in their computational state. We then use these same mechanisms to construct explanations. Specifically, we use "higher-order" applications of PP to build an *abstract model* of the world, based on concepts, from the models represented in the basic computational state. The abstract model acts as a "cache" of building blocks for explanations (a richer form of what we previously called the "external representation"), so that these do not need to be built from scratch each time.

I will refer to these elaborations of the agent's basic pre-existing computational mechanisms in Figure 3, which now include construction of world models using PP, and specialized "units" that perform automated calculations on built-in models (such as those for navigating 3D space), as its "Lower System." And I will refer to the computational state of this system, which contains representations of the Lower or "Concrete Model" as the "Lower State." Similarly, I will refer to the new mechanism that uses PP to build concept-based models of the Lower State as the "Upper System" and to its computational state, which includes the Upper or "Abstract Model," as the "Upper State." Figure 4 portrays this implementation of our mechanism for shared intentionality.

[14] The advantage of generative models is that they reason forwards, or causally, from (models of) the world to predicted observations; this can take account of sensor defects and observer behavior (e.g., displacement of sensors due to movement) and is more straightforward, simpler, and generally more accurate than the reverse inference. Most of the recent advances in AI, such as the Large Language Models employed with ChatGPT (where the "G" stands for "generative") and similar systems use generative methods. Notice that predictions are not necessarily at the sensor level (e.g., individual pixels) but can target some limited bottom-up interpretation of these (e.g., detected objects in a self-driving car). Also note that bottom-up interpretation may be used to create an initial model: this is simply the prediction error when there is no prediction.

[15] Conceptually, this is similar to a Kalman Filter, applied to complex models.

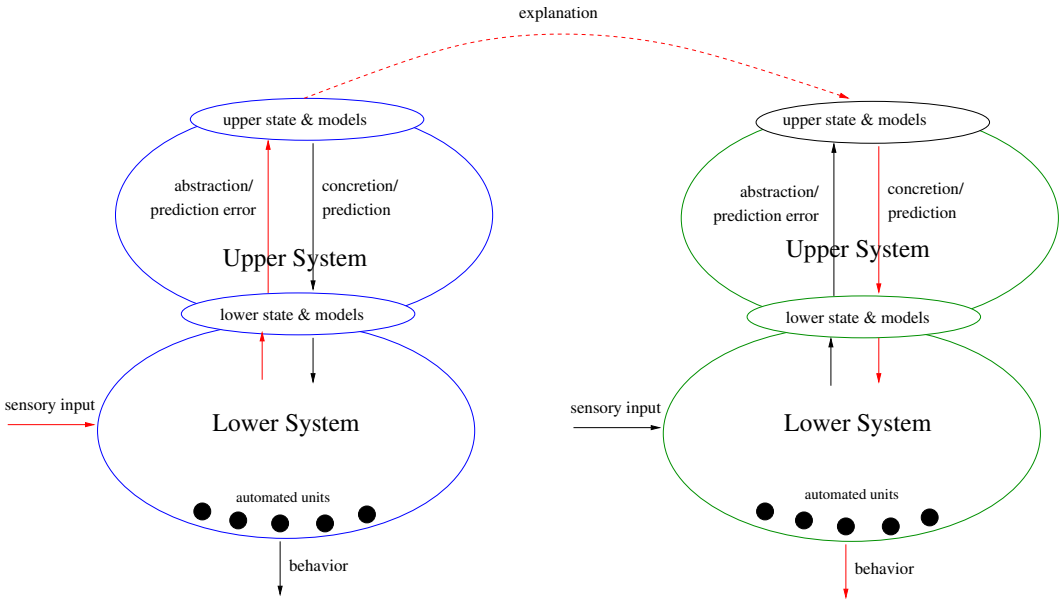

Figure 4: Architecture of Implementation to Create Shared Intentionality

Whereas previously Alice repeatedly had to build and optimize new explanations for each communication to Bob, these can now be constructed as augmentations to her persistently maintained abstract model. The "deliberative and controlled" aspect of her theory of mind for Bob will be part of her abstract model, so that rather than a separate parameter to her concretion function as supposed earlier, it simply participates in predictions along with other relevant parts of that model and, by minimizing prediction error, PP will solve the constraint previously represented as (3) and thereby generate a suitable explanation.

When Bob receives an explanation from Alice, it will first be detected by his sensors and Lower System. However, the Lower System will be unable to interpret its symbolic content and will simply add it to the Lower State. His Upper System will not have predicted this addition to the Lower State and learns about it as a prediction error. The Upper System will then extract and interpret the concepts contained therein, which will enrich its abstract model. PP then sends a concretion of this to the Lower System, where it will generate a large prediction error (since it presumably contains information that is new to Bob). Prediction errors can elicit two responses: one is to send the error to the Upper System, where it may cause revision to the abstract model maintained there; the other is to change the Lower State and its concrete model in a way that reduces the error. This cannot be done arbitrarily since the Lower State and model must be consistent with sense data from the environment. One option is to adjust the interpretation of sense data (as

when we resolve an optical illusion), and the other is to perform some behavior that will adjust the environment. William James was the first to suggest that behavior is driven by (what we call) the Lower State: "every mental representation of a movement awakens to some degree the actual movement which is its object" [78]. Thus, for example, an explanation may indicate that Bob's right hand should grasp part of a specific tree. After concretion, this will be represented as a configuration of Bob's concrete model that differs from its current state and this prediction error can be resolved by Bob actually moving his right hand and grasping the tree in the manner indicated [50, 51].

We have not said much about the Lower System, but in a typical robot it will contain a collection of automated modeling, planning, calculation, and execution "units" for specific tasks, often interacting through a "blackboard architecture" [109]. The idea here is that elements of the Lower State are deposited in a common memory or workspace (the "blackboard"), from where they are removed by those units that recognize and "know" how to interpret them: these units will produce results that are deposited back in the blackboard for consideration by other units, and/or they may generate behavior as described above.

This concludes my account of mechanisms for constructing shared intentionality among computational agents. I now claim that the mechanisms assumed and developed here are consistent with those of the human brain and can explain the emergence of shared intentionality in humans, as described in the following section.

3 Human Interpretation

The mechanisms I have proposed to endow the robots Alice and Bob with shared intentionality are plausible but probably not those that a robot designer would use to deliver such capability today: there are more powerful technologies that can create shared intentionality in robots by direct communication of models, goals, and plans in some pre-arranged shared format (i.e., something closer to a language). Nonetheless, the proposed mechanisms are perfectly feasible and I chose them because they are prelinguistic and based on capabilities known or generally considered to be present in the human brain: specifically, predictive processing, a dual-process architecture with powerful and autonomous low-level automation, and some form of global workspace. Thus, I propose that aspects of shared intentionality (or, more generally, "collective" intentionality [129]) in humans are created by the same mechanisms as those described for Alice and Bob. Of course, this assumes that evolution provided some of our ancestors (I will call them "proto-humans") with capabilities similar to those of Alice and Bob, but not yet with anything more powerful for the direct construction of shared intentionality, like language. I think this is plausible, because (temporally adopting teleological usage) there is no reason for language to have evolved prior to the construction of shared intentionality.

I will say more about evolutionary plausibility in Section 3.4 when I have described additional capabilities that I believe are associated with shared intentionality. I will argue that these related capabilities build on the mechanisms for shared intentionality and therefore it was the first to emerge: hence, I call this the Shared Intentionality First Theory (SIFT) for emergence of these capabilities.

The mechanisms developed for shared intentionality in Alice and Bob make extensive use of predictive processing (PP). This is a popular theory of brain operation [145], where it is also known as "predictive coding" [28], "predictive error minimization" [69], and (using terminology derived from statistical physics) the "free energy principle" [51].[16] Its use in computer science is to a large extent inspired by these biological precursors.

Recognition that sense interpretation must work "top down" rather than "bottom up," as conventionally assumed, was first documented by Helmholtz in the 1860s [143] and developed in more detail by Gregory [64] (who, in the 1980s, explicitly related perception to hypothesis testing in science), and by Rao and Ballard [119] in the 1990s. PP posits that the brain builds probabilistic models of its environment and uses these to predict its sensory input. The predictions are compared to sensed reality and the differences are used to refine the models via (an approximation to) Bayesian variational inference in a way that minimizes prediction error. This minimization can be achieved by refining either the upper or the lower model, or by changing the environment; the latter may be achieved by using our body to perform actions [50], as described earlier.

PP has Bayesian priors flowing from models down to sense organs as predictions and observations flowing back up as prediction errors, and this explains the otherwise puzzling fact [60] that there are many more neural pathways going from upper to lower levels of the brain than vice versa (predictions require more bandwidth than errors). PP in humans differs from that described for Alice and Bob in that the human perception system maintains many levels of intermediate models, each contributing a small step to the overall interpretation (e.g., edge and motion detection for vision [119], different time scales for speech [24]), whereas Alice and Bob's perception systems have just the single lower level. However, their sensor interpretation at that lower level will be implemented by deep neural nets whose layers will each build representations that could be regarded as intermediate models—but note that these representations are refined only during training and are fixed thereafter, whereas intermediate human models are under constant revision.

Beyond the multiple models of its perception system, the human brain is postulated to have a "dual-process" macro-scale architecture comprising "System 1" (fast, automatic, prone to error) and "System 2" (slower, requires direction and

[16] Some authors (e.g., [37, 70]) treat PP as a theory of consciousness in itself, whereas I regard it as a basic mechanism of perception that is probably present in all animals with a nervous system and brain.

effort, can perform reasoning) [42,48], popularized as "thinking, fast and slow" [82]. Note that this is a logical architecture; it need not be realized as physically separate parts of the brain. The perception system's lower-level models are found in System 1 and higher-level ones in System 2. Alice and Bob likewise have Lower and Upper Systems with corresponding concrete and abstract models. In the following, I will suggest how rationality emerges within these dual-process architectures,

3.1 Local Impact of Shared Intentionality Mechanism: Rationality

My earlier account of predictive processing was incomplete in that I described prediction errors leading to incremental model refinement. In fact, this occurs only with "small" prediction errors; "large" prediction errors indicate a "surprise," meaning the world has not evolved as expected, or the performance of our sensors has changed or failed (e.g., dazzled by the sun). Surprise causes a more drastic reappraisal of models and plans; the extent and effectiveness of the reappraisal obviously depends on the "reasoning" capability available. The mechanisms I have proposed for shared intentionality have endowed our proto-humans with an Upper System that maintains an abstract Upper Model over which they are able to construct and manipulate explanations. These capabilities provide our proto-humans with a sophisticated response to surprise and also form a foundation for independent reasoning.

In particular, just as Alice can communicate with Bob to create shared intentionality, so she can also communicate or "talk" (wordlessly) to herself. Her faculties for abstraction, explanation, and concretion can be employed in a local loop, interacting with the built-in automation provided by specialized units in her Lower System. This is portrayed in Figure 5, although in reality the self-communication is internal.

The power of this architecture is that the Upper System has the ability to construct and deconstruct explanations: that is, it can manipulate relations among concepts.[17] Freed from the need to construct a communication for Bob, Alice may be able to exploit and control the search for explanations in new ways. For example, by manipulating her model of his concretion, she can perform counterfactual and "what if" reasoning.

I propose that this local loop is the basis for human rationality, by which I mean the construction of plans and actions that may be expected to achieve their objectives despite an uncertain world. In operation, the Upper System might construct a conceptual goal; its local capabilities for manipulating concepts may rearrange or decompose some elements of this and concretion can send some of them to specialized units in the Lower System, where they will be transformed in other ways and

[17] We must be careful to postulate only limited ability here; otherwise we are invoking something close to language. What I have in mind is the "protolanguage" of Bickerton [9] (i.e., language lacking grammar) or the elementary "Isolating-Monocategorial-Associational" (IMA) grammar of Gil [56]. Young children, and pidgin speakers, are able to accomplish quite a lot with just such primitive combinations of words.

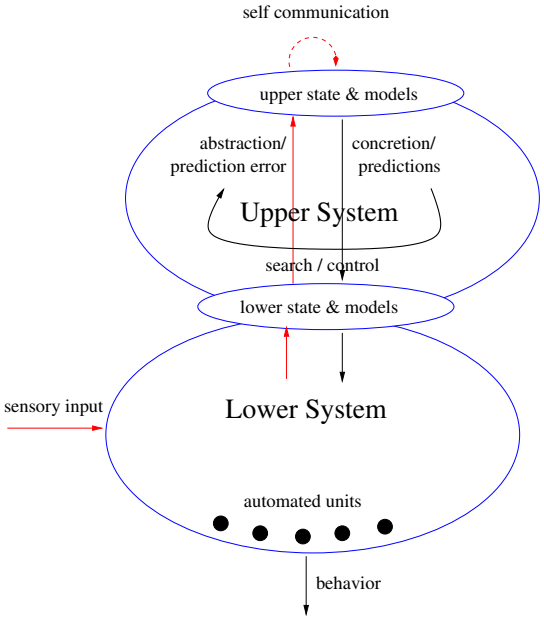

Figure 5: The "Local Loop" for Rationality

then abstracted back to the Upper System where they might be further manipulated to yield an explanation that solves the original goal (which can be checked by concretizing both goal and explanation and applying them to models of the world maintained by the Lower System). Alternatively, the goal might originate in the Lower System, but the loop will operate in a similar way following an initial abstraction to the Upper System.

The local loop is not just a problem solver: it can produce action. When Alice receives a communication from Bob, its concretion augments her Lower State and thereby influences her future behavior (recall earlier discussion that divergence between the sensed environment and Lower Model may be resolved by actions that change the environment). And this will be true of Alice's rational deliberations, too: the local loop can construct a solution to an upper- or lower-level goal and its concretion will set the Lower System on course to deliver suitable behavior.

As noted, concretion allows the Upper System to recruit specialized Lower System units as subroutines. Modern computerized reasoning systems do something similar, using highly efficient units such as "SMT Solvers" as subroutines [31]. The native problem is first transformed into a "Satisfiability Modulo Theories" (SMT) problem, solved by the SMT Solver, and then the result is translated back to the terms of the original problem. The Upper System of proto-humans can likewise use a Lower System unit that models vertical space as a subroutine in reasoning

about social hierarchy.[18] When expressed linguistically, the transformations become metaphors: "he is my superior and she is my peer, but I am above the others." Lakoff and Johnson [89] observe that this application of metaphor pervades our thinking: it is a primary mechanism of thought, not just a feature of language, and now we can see why (and that it is prelinguistic).[19]

I think this proposed architecture also illuminates the "fast and slow" dual-process model [82] and suggests why deliberative thinking is slow and easily fatigued: the "fast" mechanisms use the specialized units built-in to the Lower System that operate "in parallel," whereas the "slow" ones employ rather costly local loops, optimizing searches, and transformations that operate as a "single thread." As Kahneman illustrates, and perhaps due to its costs, this architecture for rational deliberation does not guarantee good results: the Lower System automation is often "rough and ready" and the Upper System might use it poorly via an inappropriate metaphor; furthermore, its consideration of alternative and unfavorable scenarios may be optimistic and cursory. Thus, it requires effort to derive full benefit from this capability. That effort is composed of attention and control, which are the foundations for intentional consciousness, as we will now consider.

3.2 Intentional Consciousness

Consciousness is a notoriously difficult topic, with many facets. However, it is generally agreed that two of these are primary: *intentional* consciousness,[20] and *phenomenal* consciousness. The first concerns the ability to direct attention and to think *about* something and to know that you are doing so. The second concerns "what it's like" to have subjective experiences such as the smell of a rose or a feeling of pain, experiences referred to generically as *qualia*.

We started this investigation by considering mechanisms that can construct shared intentionality, so it will not be surprising if these mechanisms also deliver individual intentionality and, thereby, this facet of consciousness. Let us consider how this might come about.

When a proto-human Alice constructs an explanation to communicate to Bob, or engages in rational deliberation using a local loop, she surely needs to focus her resources on these tasks and on the explanations generated and received. The lower

[18] The Lower System or subconscious in humans is thought to have numerous specialized faculties, although their identity is open to debate. Evolutionary psychology might argue for mental "modules" such as those for social rules, mate selection, and so on [113], while embodied cognition might argue that specialized calculation derives from our interaction with the physical world, and concerns reasoning about distance, weight, time, etc. [81]. For my purposes, it is sufficient to acknowledge the existence of specialized cognitive units, without worrying about their precise functions.

[19] Lakoff [88] describes mental mechanisms for metaphor that are more complex than mine, but they are not incompatible.

[20] Also called *access* or *cognitive* consciousness.

or subconscious system state records and represents a vast amount of information from all our sense organs, models of the world, memories, beliefs, desires, intentions, and all the ongoing automated processing of these. The abstraction mechanism must be highly selective about what of this it chooses to represent in its upper model and to use in constructing its (one) current explanation: this selectivity and focus is what we mean by *attention* and my proposal is that intentional consciousness corresponds to *awareness* of attention and of the things attended to in the Upper System. In fact, I suggest that intentional consciousness goes beyond awareness of attention and also has elements of executive control, so that we can focus our resources on particular deliberations (unfocused deliberations will interfere with each other and make little progress). This focused deliberation seems different than the automated mental processes of the Lower System, which operates its units "in parallel" (i.e., simultaneously), weighted and directed according to the exigencies of the moment. For our Upper System deliberations, we are largely able (and required) to maintain a directed focus that seems, despite occasionally "wandering," to operate as a single "sequential" thread that is "about" some topic or goal.

This single-threaded purposeful focus of attention and control, and awareness of it, seem sufficiently different than other mental processes that it should not be surprising that it corresponds to a unique mental experience which, I propose, is intentional consciousness.[21] However, although this explains what the computational processes underlying intentional consciousness are for and how they are constructed, I cannot explain how they produce an apparently nonphysical experience.[22] This is a problem for all materialist theories of consciousness and, unlike Graziano *et al.* [63] and others, such as the "illusionist" school [49, whole issue], I do not think the problem is solved by claiming that these processes cause the brain merely to *report* or to register *belief* in a nonphysical experience. Thus, I recognize that a leap of faith is required to accept this interpretation of intentional consciousness, but I argue that it accords with observation, and with some other theories of consciousness.

First, I need to expand our previous considerations of shared intentionality and rationality, where I implicitly supposed that the mechanisms of the Upper System (i.e., abstraction, concretion, and manipulation of explanations) are invoked periodically to accomplish specific acts of communication or rational deliberation. Now I suggest that these mechanisms do not otherwise sit idle, but operate continuously and autonomously and we attend to them deliberately, and are conscious of them, only when focus is necessary. As life proceeds, the subconscious Lower System gen-

[21] This contrasts with Dennett's "multiple drafts" theory [34]: I accept there may be many fleeting drafts or models at the subconscious level, but consciousness resides with the one explanation that is currently a candidate for communication or the subject of local deliberation.

[22] Observe, too, that in constructing an explanation for Bob, Alice surely needs to recognize herself as distinct from him. Hence, I speculate, but once again cannot explain, that these processes also produce a "self-model" [103] and the illusive experience of personal identity [74, Book I.iv, section 6].

erates behavior, sometimes at the behest of the Upper System but often on its own. This is constantly monitored by the Upper System, which maintains an abstract model and generates explanations whose concretion closely tracks the Lower System model and state. Although our behavior is mostly generated by the Lower System, we are conscious only of the abstractions and explanations reconstructed by the Upper System; thus, it "feels" as if consciousness causes behavior.

This explains some otherwise puzzling facts. For example, experiments such as Libet's [92], and studies with "split-brain" patients [54], reveal that, contrary to our intuitions, the conscious mind is less an initiator of actions and more a reporter and interpreter of actions and decisions initiated in the subconscious. But we can now see that the primary purpose of the mental faculty that supports consciousness is precisely the reporting and interpretation needed to construct shared intentionality; rationality and consciousness ride on the mechanisms of shared intentionality and share its character.

Next, the experience of intentional consciousness is derived from the Upper System, whose foundational purpose is to construct communicable prelinguistic abstractions that can deliver shared intentionality. This explains why much of the experience of intentional consciousness is of an inner dialog (rather than, say, a stream of images). Initially the dialog would be wordless because it is difficult to imagine how speech could have evolved prior to shared intentionality, but language, and ultimately spoken language, could surely have developed quite rapidly once shared intentionality became available: the "gist" that is at the heart of language understanding and memory [14] could have evolved from the wordless concept-based explanations of shared intentionality, which might also be a basis for "mentalese," the "language of thought" [121].

To consider further observations that can be explained by this theory of intentional consciousness, we need to see how phenomenal consciousness fits in.

3.3 Phenomenal Consciousness

When Alice explains her idea for a log-bridge to Bob, she might finish by pointing to a particular log as the one to be used. This is a remarkable thing: she is making an external reference to a subjective inner experience, namely her visual field.

Once we have the ability to construct shared intentionality, we need the additional ability to communicate our perceptions of things in the world and things about ourselves. We experience these through our senses and so to reference them in communications it is necessary for subconscious information derived from our senses to be abstracted into the Upper System.

I submit that this is phenomenal consciousness: in order to communicate the things we sense and feel, selected abstractions and explanations about them must be present in the Upper System and its model—and we will be conscious of them just

as we are conscious of other content in our abstractions and explanations. It is one thing for our visual system to allow us to sense a log that we may choose to sit down on—all this can be done subconsciously and "in the dark," as it is by people with "blindsight" [35, 75]—and quite a different thing for our phenomenal consciousness to present us with the experience of the visual field, so that we can indicate "the log on the left." Notice that this indication is symbolic, it is not a direct response to the sensation concerned, as when we sit down on a log or recoil from pain. And notice, too, that consciousness of the relevant sense arises only because we may need to communicate it to others.

We have senses beyond the classic five, possibly as many as 50 [144]; it is notable that we are conscious of some of these—and to different degrees—and others not at all. For example, the sense of balance (kinesthesia) is important and we are conscious of being in or out of balance, but we are not conscious of the sense organs (e.g., semicircular canals) that support kinesthesia (except their *mal*function, as when we feel dizzy). And for another: I have a colleague whose proprioception is failing and he finds it difficult to describe the symptoms because this sense is largely unconscious and consequently we have no vocabulary for "what it's like."

It seems that the senses of which we are phenomenally conscious are just those that it can be useful to communicate explicitly to others, notably including the classic five. We are not phenomenally conscious of the senses supporting kinesthesia because we do not need to communicate their detailed content (e.g., "the current rate of yaw is 5 degrees per second"). So there is little "that it's like" to experience the senses supporting kinesthesia—and nothing at all for proprioception. Hence, I think it is quite possible that there is nothing "that it's like" to be a bat [107] because, assuming bats do not create shared intentionality, they lack an Upper System and all their senses and actions operate nonconsciously, "in the dark."

In addition to our senses, we need to communicate certain subjective experiences such as hunger, thirst, and pain (as a sensor for injury or sickness), together with moods and emotions, so these must be abstracted to the Upper System as well. It does not particularly matter how the color of a red rose is represented in the Upper System, nor the anguish of jealousy (i.e., "what it's like" to experience these), as long as they can be distinguished.

This proposal partially solves the "hard problem" of consciousness [25]: why are some phenomenal states conscious? They are conscious because we need to communicate them to others and so their abstractions must partake in explanations and be present in the Upper Model, attended aspects of which are conscious. I say "partially solves" because it explains what the process underlying phenomenal consciousness is for, how it works, and why it produces a nonphysical experience—but it does so by establishing a relation to intentional consciousness and, although I have also explained what that aspect of consciousness is for and how it is constructed, I accept that I cannot explain why or how it produces a subjective experience.

Furthermore, this proposal does not explain why our subjective experience of, say, red is what it is and not something else. I am unapologetic about this: my opinion is that it is an unanswerable question, and an unimportant one: as stated above, red has to have some representation in the Upper System so that we can reference it ("the mailbox is the red thing"), but the form of that representation and "what it's like" do not matter, provided they are distinct from those of other perceptions. Neither does it explain "what it's like" for different people to experience similar qualia: is your perception of red the same as mine? There seems to be no requirement for this to be so and the existence of synesthesia [65] or the phenomenon of "The Dress" [147] indicate that people can indeed experience the same qualia differently.[23]

Many will find this disappointing: phenomenal consciousness is the experience most central to our lives. But the fact that its representation is arbitrary does not diminish its significance. What matters is that we *care* about the form that it does take, an attribute sometimes referred to as *sentience* [40].

A less discussed variant on this question asks why our phenomenal consciousness occurs at the representational level that it does [77]. Vision, for example, employs dozens of specialized units that build representations ranging from a "primary sketch" at the lower levels (which are unconscious), through the shaded, 2.5D, photograph-like "intermediate level" (where my phenomenal consciousness currently sees a black parallelogram on top of a partially occluded larger brown parallelogram) to a top-level conceptual model (where I recognize my keyboard on top of a cluttered wooden desk). Jackendoff [77] observed that the intermediate level (as identified by the theories of sensory interpretation current at that time [97]) seem to be where phenomenal consciousness is located. Prinz [117] argues that this observation remains true under modern theories, and the question is why is it this level, and not some higher or lower one?

Marchi and Hohwy attempt to answer this using a PP model of brain operation [94]. They argue that it depends "on the spatiotemporal resolution of the typical actions that an organism can normally perform." For humans, this makes the intermediate level appropriate, but this is "not an essential feature of consciousness; in organisms with different action dispositions the privileged level or levels may differ as well."

I speculate that SIFT provides a simpler explanation: the representational level of consciousness is whatever was the top level prior to the evolution of the Upper System. In Figures 4 and 5, the Upper System abstracts from and concretizes to what is labeled as the "Lower State & Models." My speculation is that these models correspond to the (collection of) top-level representations prior to the evolution of the Upper System. These representations are not themselves conscious; phenomenal

[23] The Berlin-Kay theory [8], whereby the selection of basic color terms in a culture is predicted by the number of such terms, suggests to me that mostly we do experience qualia similarly.

consciousness resides in the Upper System's higher-level representations of (or references to) these. Viewed relative to the high-level conceptual models built by the Upper System, the target of phenomenal consciousness is indeed an intermediate level, but that is a consequence of the subsequent evolution of the levels above it.

3.4 Biological & Evolutionary Plausibility and Evidence

I have advanced a proposal for shared intentionality, rationality, and consciousness that I call the "Shared Intentionality First Theory" or SIFT. The proposal is that these mental attributes form a "package" but that shared intentionality provides the framework on which the others are constructed. The proposal is based on computational constructions for hypothesized humanoid agents or robots. However, the constructions assume an underlying computational architecture with capabilities selected from those known or believed to exist in humans. These include predictive processing and a dual-process architecture with multiple specialized "units" for automated lower-level calculation, likely organized around a global workspace.

This proposal is entirely abstract and computational: I assume that the purpose and function of the human brain is to perform calculations, notably those concerned with construction and interpretation of models of its environment. I join others [16] in maintaining that progress in understanding emergent properties such as consciousness requires abstraction and theories in addition to experimental neuroscience and, eventually, development of bridge laws between these points of view [87]. As yet, I cannot identify biological mechanisms or structures or "neural correlates" that correspond or bridge to my overall proposal,[24] but there are papers that do so for its constituent parts (e.g., [70], [42, section on Neuroscientific Evidence]). Thus, I posit that my proposal is biologically plausible. I further posit that it is plausible that the overall package of capabilities evolved from these constituents.

Humans have shared intentionality whereas other primates do not [18], so this capability emerged sometime in our recent evolutionary history and we can ask whether it emerged before, after, or with related capabilities such as language, rationality, and consciousness. The basis of SIFT is that the mechanisms of shared intentionality evolved first, or provided the survival and reproductive advantage that caused natural selection to favor the package. This differs from other theories of rationality and consciousness, which go wrong, in my view, at their first step: they assume, as is natural, that since consciousness is subjective and personal, it must do something for the individual. This is not to deny that consciousness and rational deliberation have benefit for the individual, only that their evolutionary origin lies in shared intentionality leading to teamwork that delivers advantage to the group [4]. I am aware that group selection is a contested notion; however, once this

[24] Recent papers that look at neural correlates of shared intentionality (e.g., [45]) focus on the synchronization or "coupling" of minds rather than active communication between them.

package of capabilities exists, evolutionary selection can operate on its components for individual advantage.

A key requirement of SIFT is evolution of an Upper System that abstracts the Lower State in terms of concepts and can manipulate those concepts to form explanations. Some will find circularity in this invocation of concepts: to communicate we need a shared abstract vocabulary, which I associate with concepts, but how did these arise without communication?

As we have noted before, it is generally understood that the function of the brain is to guide an animal's interaction with its environment [28] and to do this it must build models of that environment [29]. The models cannot be in terms of raw sense data; even simple animals must perform some categorization on that data: they must surely distinguish rocks from plants, and plants from animals, and their own species from prey and predators. Zentall *et al.* [150] provide a survey of concept learning in animals and conclude that similar underlying processes apply to humans. Carey [19] understands human concept formation to occur on two levels; core concepts (which are all we require here) are acquired rapidly and early in childhood by the processes mentioned above, and language is required only for higher-level concepts, such as "The United Nations." Thus, human core concept formation could have achieved detailed categorization of the natural and social world prior to development of the mechanisms of communication developed here. However, for communication, and the construction of shared intentionality, it is not enough to have concepts: they must be held in common; when Alice points to a tree, Bob must think "tree" not "leaves." Fortunately, there do seem to be prelinguistic mechanisms that ensure core concepts are shared among members of a local community [133, 142].

Thus, I maintain it is feasible that the Upper System evolved to perform the functions of shared intentionality—but it is also possible that these functions were adapted from some prior dual-process architecture that evolved for other reasons (e.g., to manage "surprise"). Either arrangement suits my purpose (though the former suggests that the Upper System is unique to humans while the latter does not). Cognitive structures such as these leave no physical evidence in the fossil record so we must look to archaeological and anthropological evidence of behavior to see if shared intentionality did emerge first among the package that includes rationality and consciousness.

It seems there is evidence for collective hunting by ancestral humans going back millions of years [101], but it is not clear whether this indicates shared intentionality or merely a built-in program for group behavior like that of wolves. More definite signs of shared intentionality, such as living in large groups,[25] are seen in modern and possibly archaic humans dating back a few hundred thousand years [111, 138].

[25] In small groups, everyone knows everyone else and some form of group behavior can develop based on individual and collective relationships; in large groups, we need rules, and these need shared intentionality.

For signs of rationality and consciousness, I believe we have to look much later, to the "explosion" of creativity (cave paintings, hand prints etc.) seen in the human record about 40,000 years ago.[26] The archaeologist Steven Mithen attributes this new behavior to integration of formerly separate cognitive domains [105] and this would be consistent with emergence of the "local-loop" mechanism for rationality described in Section 3.1 that is able to exploit multiple automated Lower System "units" (via metaphors). It is contested whether Neanderthals, who became extinct soon after this date, engaged in symbolic thought [134], and consideration of how their behavior and mental attributes differed from modern humans would be interesting from a SIFT perspective. There is clearly opportunity for more inquiry and evaluation of evidence here, but it does seem that the human evolutionary and archaeological record may support, and certainly does not contradict, the theory of Shared Intentionality First.

Another way to seek evolutionary evidence would be to look for precursors to shared intentionality, rationality, and consciousness among living species [41]. Some attribute shared intentionality to social animals such as wolves, and even bees, whereas others claim their collective behavior emerges from simple rules, preprogrammed by evolution [38]. I join with the latter and believe that one individual of these species cannot communicate a new idea or plan to another, save by imitation.[27] Other animals, even primates, show few signs of shared intentionality [18,38,59] but some are popularly believed to be conscious [3,122], which would contradict SIFT. However, we know from Libet's experiment [92] and its successors that humans attribute behaviors to intentional consciousness that actually originate in the subconscious. I suspect that many of our behaviors are like this, and that when we see similar behaviors in animals, we attribute them to consciousness because we falsely believe that is how it is with us. Thus, although there are opportunities here for tests and possible refutations of SIFT, my belief is that consciousness evolved sufficiently recently that it is not to be found among our ancestor and sister species.[28]

Hence, I suggest it may be more productive to look for SIFT-like developments through convergent evolution among hypersocial species such as elephants, toothed whales, and corvids. Certainly, sentience and possibly consciousness are sometimes

[26] These records were first found in Europe and dated to around 20,000 years ago; more recent investigations have found precursors in Indonesia dated to 40,000 years ago and in Africa to as long as 100,000 years ago. All of these are later than emergence of shared intentionality.

[27] Domesticated dogs are an interesting case because it is possible they can participate with humans in "asymmetric" shared intentionality: that is, humans may be able to communicate a goal or plan to dogs, but not vice versa. There is some evidence that dogs understand human intentionality [127], so it is possible that a symbolic utterance such as "fetch" (the ball) is processed this way, but it could also be a conditioned reflex. Furthermore, dogs have been bred selectively by humans for thousands of generations, so it is possible that their mental faculties have been selected along with their appearance and behavior, and do not represent the capabilities of dogs in the wild.

[28] Blindsight in apes does provide a possible contradiction to SIFT, since it suggests that normally sighted individuals are conscious of their visual field [75].

attributed to these [122] and it might be enlightening to investigate their capacity for shared intentionality. Octopuses are another interesting and challenging case, as they are widely thought to be intelligent and possibly sentient, yet they are not social [57]. However, each arm of an octopus is capable of autonomous behavior and has a concentration of neurons somewhat like a brain; together, these contain twice the neurons of the main brain (put another way, each "arm brain" is a quarter the size of the main brain) [21]. The neural pathways from the main brain to those in the arms are too small for high-bandwidth integration, so it is possible that a single octopus instead creates shared intentionality (possibly of the asymmetric variety discussed previously with regard to humans and dogs) among the "community" of its nine separate "brains" [29] and that SIFT then delivers more advanced capabilities, possibly including consciousness (which could be very different to that of humans—lacking unity [20, 99], for example). There may be opportunities for research here.

It is shared intentionality and intentional consciousness that create a rôle for phenomenal consciousness; thus SIFT predicts that phenomenal consciousness evolved with or later than shared intentionality, and that would imply it is unique to humans (plus possibly those animal candidates for shared intentionality mentioned above). This is contradicted by those who believe it is part of the basic mechanism of advanced perception and arose 500 million years ago (in the Cambrian explosion) and is possessed by all vertebrates [44]. Evaluating these competing theories is complicated by lack of any accepted means for assessing phenomenal consciousness in animals. Of course, it may be that precursors to the components of SIFT evolved at different times and in different orders to emergence of the finished package, and there may be opportunities for falsifiable experiments here.

Independently of evolution, we could look for direct evidence in modern humans for some of the mechanisms I have hypothesized. For example, infants can provide an opportunity to evaluate shared intentionality versus consciousness in prelinguistic humans [106] and the impact of shared intentionality on cognitive development [139]. These investigations must be driven by precise hypotheses and, since my proposal is new, much of it remains work for the future. However, at least one relevant capability has been observed in adult humans: this is the "interpreter module" identified by Gazzaniga [54, Chapter 3]. This module selectively attends to what is going on elsewhere in the brain and retrospectively constructs explanations for the beliefs and behavior produced. This is like a version of the abstraction and explanation capability that I have hypothesized. One possibility is that the interpreter module *is* this hypothesized capability; another is that it evolved separately—but it is difficult to see its utility prior to shared intentionality, and it would be redundant afterward. There are opportunities for further investigation here.

[29] Jennifer Mather disputes this and likens the arms to "subroutines" of the main brain [100, figure 1], but I think there remains the question of how the "remote procedure call" is communicated.

4 Comparison with Other Theories of Consciousness

There are many theories of consciousness (of which I focus on the materialist variety): some such as Neurobiological Naturalism (NN) posit ancient evolutionary origins [44]; others, such as Global Workspace Theories (GWT) [5, 32], focus on biological processes; some, such as Integrated Information Theory (IIT) [140], Orch-OR [67], and Panpsychism [58], favor physical explanations and mechanisms; yet others, such as Higher Order Thought (HOT) [17, 55, 123] and Attention Schema Theory (AST) [60] hypothesize architectural "dual-process" structures in the brain [42, 48, 82]. Graziano *et al.* [63] compare and reconcile several of these theories with AST, and the comparisons remain largely valid with SIFT substituted for AST. However, none of these other theories claim to explain what consciousness—and phenomenal consciousness in particular—*does*, nor what it is *for*.

SIFT is different in that it focuses on specific purposes to be accomplished, and develops mechanisms to achieve these, starting with shared intentionality (to achieve teamwork) and proceeding to rationality, which is seen as a fortuitous side-effect, built on shared intentionality: "teamwork for one".[30] Intentional consciousness is then identified with awareness of attention to, and control of, the mechanisms of shared intentionality and its Upper State and models. Phenomenal consciousness arises because we need to communicate aspects of our sense experience and subconscious Lower State: hence these must be abstracted into the Upper State of shared intentionality, of which we are conscious.

My presentation of proposed mechanisms provides context for several of the theories identified above; in particular, it delivers a "dual process" architecture that is consistent with, and explains some aspects of, existing dual-process models of brain function [42, 48, 82]. Dual-process models hypothesize a logical, not physical, organization of the brain, and it is quite possible that they are realized by physical and neuronal mechanisms with quite a different structure. Plausible candidates include global workspace theory [5] and global neuronal workspace [32].

The Upper System in my dual-process model, performs calculations whose inputs and outputs are subconscious mental states located in the Lower System: it is a part of the brain that senses and writes to other parts of the brain. While the subconscious Lower System builds representations and models of the external world, the Upper System builds representations of those representations; these constitute

[30] For example, at some time, we have surely all said "I cannot explain it, but I can show you how to do it." Elsewhere, we suggest how a task description can be inferred from demonstrations by inverse reinforcement learning [79]; thus, Alice can construct an abstract model of some task that her Lower System "knows" how to do by mentally demonstrating it to herself.

what computer scientists call a "reflective system" [148] and what philosophers refer to as "Higher-Order Thought": that is, thoughts about thoughts [55, 124].[31]

Dual-process and HOT theories generally hypothesize some process whereby activity in the subconscious results in awareness at an upper or higher-order level but do not describe what purpose this serves. Unlike most of these theories, SIFT explains what the upper system does (and hence why it might have evolved). Furthermore, as Graziano observes [60], higher-order awareness must also affect the subconscious lower-level activity (otherwise it is impotent), and few theories address this. In SIFT, these two directions respectively correspond to abstraction and concretion of explanations that are mechanized by the well-accepted operations of predictive processing, and both are involved and coupled in the construction of upper level models and explanations, and lower level behavior.[32]

Graziano's AST [60] and some HOT theories [17] are the ones closest to SIFT: in our terminology, Graziano proposes that the brain constructs a model (he calls it a schema) of the targets of attention and it is "aware" of this model and that awareness constitutes intentional consciousness.[33] SIFT proposes that our Upper System builds models of the Lower System and these support the construction of explanations; intentional consciousness then corresponds to awareness of the single-threaded process of attention and control that manages and applies the resources of the Upper System to these tasks. Graziano has to postulate the attention schema because he does not otherwise have an upper-level system that can provide a location for consciousness, whereas SIFT has its Upper System, so that consciousness can reside directly with its processes of attention and control.

This implies that consciousness in SIFT applies to a specific locus of attention and control—that concerning the Upper System and its construction of Upper Models and explanations—whereas AST applies it more generally. For example, Graziano believes that higher animals are aware (i.e., conscious) of other animals' focus of attention (e.g., "he is looking at me") [62], whereas I consider this level of (mutual) attention could be unconscious and would become conscious only if elevated into an Upper System and formulated as the *explicit* explanation or thought "he is looking at me."

Consciousness in SIFT is awareness of attention and control in the Upper System; some experiments are claimed to demonstrate that attention and awareness are different [90] and likewise attention and consciousness [85]. However, when I say "con-

[31] This is among the oldest conceptions of consciousness, dating back at least to Locke in 1689: "Consciousness is the perception of what passes in a man's own mind" [93, Book II, Chapter 1, Section 19].

[32] Oddly, most HOT theories do not relate their higher-order aspect to dual-process theories, nor do they invoke PP as a mechanism that can build the higher-order system; an exception is Lau [91], who does mention PP but opts for a first-order theory.

[33] "Awareness is a schematic, informational model of something, and attention is the thing being modeled" [61].

sciousness is awareness of attention. . . ," I am merely aligning with Graziano's usage and do not intend a specific interpretation of "awareness" distinct from consciousness; I could equally well have said "consciousness *is associated with* attention. . . " Furthermore, as noted above, this association is not with attention generally, but with attention to the Upper System and its construction of Upper Models and explanations. Thus, I interpret the experimental findings as applying to attention in the Lower System and not to that which I associate with consciousness.

SIFT delivers a very specific aspect of shared intentionality: how to communicate a goal, plan, or idea from one individual to another, prior to the evolution of language. (As noted earlier, it is difficult to see how language could have evolved prior to shared intentionality but plausible that it could do so afterward, given the mechanisms described here.) However, there are precursors to shared intentionality that provide related capabilities; these include the "social brain" (i.e., living in groups with complex social systems) [39], "cooperative communication" where individuals "align their mental states with respect to events in their shared environment" [142], "shared agency" [15], which describes the mutual plans that shared intentionality needs to bring about in order to achieve teamwork, and the general "theory of mind" [98]. These provide some of the necessary milieu for the emergence of shared intentionality as considered here, but do not substitute for it. I should also note that we have focused here on construction of a single communication within a process to achieve shared intentionality on some topic. It may take more than one communication to convey a complete idea and Jha and I suggest how inverse reinforcement learning could be used to accomplish this over a series of communications [79].

There is substantial and significant prior work on shared intentionality by Tomasello and others [136, 137, 139], but this work tends to focus on what we might call "immediate" intentionality, founded on joint attention, such as sharing information ("there is food over there") and goals ("let's go to the water hole"), and not the communication of new ideas and future plans. I suggest that shared immediate intentionality can build cooperation, but not teamwork, and it is teamwork that truly sets humans apart. Similarly, work on the origins of language suggests several sources, such as child rearing [43] or self-advertising [36] but not shared intentionality—with the exception of Bickerton [10] who relates language with the organization of scavenging, which I would again classify as shared immediate intentionality.

I have not found prior work that explores mechanisms for shared intentionality of the kind considered here, nor any that derive rationality and consciousness from shared intentionality, but there is work that relates aspects of consciousness to group communication. Frith, in an influential paper of only two pages [52], "sketches a conjecture" that consciousness enables interaction with others: "Shareable knowledge (which I equate with the contents of consciousness) is the necessary basis for

the development of language and communication. In this account, the major mistake of most theories of consciousness is to try to develop an explanation in terms of an isolated organism."

Oakley and Halligan [110] give an account similar to Frith's, but with more detail; they claim that consciousness has no executive function and is basically a "personal narrative" about processes and actions generated by nonconscious systems. Aspects of this narrative can be shared with others through "external broadcasting" and this provides evolutionary benefits. They discuss the experimental literature and cite several others (including some mentioned here) who "accept that any evolutionary advantage lies not in the 'experience of consciousness' itself, but in the ability of individuals to convey selected aspects of their private thoughts, beliefs, experiences etc. to others of their species."

Along these lines, Baumeister and Masicampo claim that "the purpose of human conscious thought is participation in social and cultural groups" [7]. They see reasoning and intentional consciousness as serving higher-level purposes that make groups more effective but do not single out shared intentionality. Humphrey also associates phenomenal consciousness (he calls it "sentience") with social purposes [76]. Similarly, de Bruin and Michael [30] suggest that Predictive Processing with upper level models informed by a theory of mind enables effective social cognition, while Sperber and Mercier [102] posit that the purpose of human reasoning is evaluation of possibly false information supplied by others; Dessalles [36] attributes similar functions to language.

All these authors accept or assume that (what I characterize as) shared intentionality is among the collection of capabilities associated with consciousness and that the collection provides evolutionary benefit. However, they seem, implicitly, to assume "consciousness first" or "reasoning first" theories rather than giving primacy to shared intentionality, and they lack models for the underlying representations, algorithms, and implementations. They assume an ability to manipulate and communicate concepts without proposing how this can be constructed on more basic foundations, and do not provide a strong path from one capability to another, nor do they explain phenomenal consciousness.

In contrast, I propose that "Shared Intentionality First" provides the most plausible basis for the construction of the package of capabilities that includes consciousness and rationality, and that consideration of the mechanisms, representations, and algorithms required for its construction leads quite naturally to the other components of the package, including phenomenal consciousness.

5 Artificial Consciousness

There is a maxim, generally attributed to Richard Feynman, that to really understand something you have to be able to recreate it.[34] Accordingly, there is a subfield of consciousness research that explores the possibility of building conscious robots, typically based on some theory of human consciousness. At the very least, these endeavors force elaboration of sufficient detail in the chosen theory that it can be simulated in a computational agent, and they also force articulation of what consciousness might be in such agents and how it can be detected.

We could apply this to the SIFT hypothesis and ask whether robots constructed along the lines described for Alice and Bob might be conscious—and if not, we could ask whether this casts doubt on the hypothesis. Before doing this, we review something of the history of attempts to develop conscious robots, also referred to as artificial or machine consciousness. This discussion is based on [125, Section 5].

Early experiments were conducted by Tihamér Nemes in the 1960s [108], but intelligence and consciousness were not sharply distinguished at that time, nor were cybernetics and (what became) AI. A modern view of robot or artificial consciousness is attributed to Igor Aleksander in 1992 [2], who postulated that such a robot would need representations for depiction, imagination, attention, planning, and emotion, and that consciousness could emerge from their interaction.

The first large project to explore artificial consciousness was CRONUS [96]. This was predicated on the idea that internal models of the system's own operation (i.e., what computer scientists call "reflection" [148], and the related philosophical notion of a "self-model" [104]) play an important part in consciousness. Physically, CRONUS was an *anthropomimetic* robot (i.e., one closely based on the human musculoskeletal system) equipped with a soft-realtime physics-based simulation of itself in its environment. The internal simulation allowed the robot to project the effects of possible future actions, which the authors describe as "functional imagination" [95]. Later studies used an even more complex robot ("ECCEROBOT"), while earlier ones had used a very simple, nonanthropomorphic device [73]. It is debatable whether complex robots added a great deal to these experiments, and they certainly increased the engineering challenges.

Like CRONUS, most recent explorations of artificial consciousness generally favor reflective architectures that employ explicit models of self. Experiments by Chella and colleagues explored such robots' interaction with others [26, 27]; here, models of self applied to "others like me" provide a theory of mind [149], and scenarios enacted by these models can be communicated to others (directly, not in the manner proposed for Alice and Bob) to create a form of shared intentionality [149]. These capabilities can be used for "inner dialog" that provides rationality in a way that resembles the local loop of Section 3.1 [73, 114];

[34] Written on his blackboard at the time of his death: "What I cannot create I do not understand."

Gamez describes other projects performed around the same time [53]. All these experiments, and those mentioned above, employ some form of reflection or HOT as their underlying theory of consciousness. Others have built systems based on GWT or IIT; Reggia provides a survey [120]. None of these projects, nor those mentioned earlier, claim to have demonstrated artificial consciousness and I suspect the same would be true of Alice and Bob, despite their design being based on mechanisms thought to correspond to those of humans.

Research on artificial consciousness seems not to have a central forum for presentation of results and discussion of ideas: the *International Journal of Machine Consciousness* began publication in 2009 but ceased in 2014.[35] Perhaps as a result, recent work seems to retread familiar ground. For example, a paper by Dehaene, Lau and Kouider from 2017 [33] presents the authors' theory of consciousness (global availability as in GWT, plus reflection built on PP), then asserts that a machine with these capabilities "would behave as though it were conscious" [33]. In a response, Carter *et al.* [23] observe that Dehaene and colleagues ask and answer the wrong questions—essentially, Dehaene *et al.* are aiming for intentional consciousness, whereas Carter *et al.* think that phenomenal consciousness is what matters: for machines to be conscious, "we must ask whether they have subjective experiences: do machines consciously perceive and sense colors, sounds, and smells?" They posit "a more pertinent question for the field might be: what would constitute successful demonstration of artificial consciousness?" This is an old question (e.g., [13]) that still seems to lack good answers.

An event in June of 2022 illustrates this: a Google engineer working with their Large Language Model (LLM), LaMDA (Language Model for Dialogue Applications), claimed it had become "sentient" and possibly conscious (Washington Post, 11 June 2022). A flurry of discussion ensued, with most commentators rejecting the claim of consciousness, but lacking a firm basis for doing so [131]. Modern LLMs easily pass standardized tests for intelligence and knowledge (e.g., high school, medical, and law examinations) and, arguably, traditional benchmarks for human-level cognition such as the Turing Test [141] (thereby motivating proposals for more demanding tests [132]). Furthermore, they can generate interactive text that speaks coherently about feelings and experiences. On the other hand, there is nothing in their internal operation that resembles any theory of consciousness, and no credible explanation how consciousness might emerge from the way they do operate.

Although deliberate and accidental research has not unequivocally demonstrated artificial consciousness, it is sharpening discussion on how consciousness could be detected, particularly since it might not resemble human consciousness (the conjecture of octopus consciousness raises the same issue). Some of these discussions distinguish *strong* and *weak* forms of artificial consciousness [71] (sometimes framed

[35] It has recently been revived as the Journal of Artificial Intelligence and Consciousness.

as duplication vs. simulation). Strong artificial consciousness would *be* conscious, whereas the weak form exhibits behaviors and attributes associated with consciousness without actually possessing it (cf. "philosophical zombies," which are considered below).

Most researchers think that simulations of theories of human consciousness can create, at best, the weak form of artificial consciousness, and that the weak form does not lead to the strong. By analogy, we can build extremely accurate simulations of the cosmos and explore the missing mass (attributed to dark matter and dark energy), yet the simulations do not *have* mass; so a simulation *of* consciousness will not *have* consciousness.

On the other hand, the weak and strong distinction seems to matter only for phenomenal consciousness: we likely will regard an entity that *has* feelings differently than one that merely simulates them. But weak intentional consciousness is operationally equivalent to the strong form: if weak intentional consciousness enables some new cognitive capabilities, then the underlying system can strongly possess these by running the weak simulation as a subroutine. This asymmetry between weak and strong forms of phenomenal and intentional consciousness is related to the "hard problem" of consciousness [25] and provides another way to formulate it.

These observations have some impact on the possibility of philosophical zombies, which are hypothetical entities built on a biological substrate that lack consciousness but reproduce levels of cognitive and social performance that are indistinguishable from conscious humans [84]. The question is whether such entities are possible. As we have postulated them, biological zombies lacking intentional consciousness would be unable to focus and control the operation of their Upper System and would be unable to communicate the content of their sense experience; they would be unable to utter truthfully such simple phrases as "I smell a rat." However, we could postulate entities with computational mechanisms that resemble our humanoid robots and thereby possess the behavioral characteristics of intentional consciousness without the experience of phenomenal consciousness. If such entities chose to fake the experience (as modern LLMs would enable them to do), they would come very close to philosophical zombies.

The general lack of success in efforts to create artificial consciousness, and its likely absence in the Alice and Bob robots, may seem to cast doubt on the mechanisms hypothesized here to create consciousness in humans. My opinion is that these judgements are premature: just as we have machines that fly but do not flap their wings, and we had to acquire deep knowledge of aerodynamics to reconcile these different approaches to flight, so our characterizations of consciousness may be too crude and coarse to identify significant nascent properties in Alice and Bob. Birch *et al.* nominate five separate elements in consciousness that should be investigated individually [11], while Holland suggests that artificial consciousness should be reframed independently of organic life forms [72].

6 Summary and Conclusions

The theory and speculation I have advanced here as "SIFT" comprise several basic claims that are somewhat independent and can be evaluated separately. The initial and central claim is "Shared Intentionality First": this is the idea that human rationality and consciousness emerge from a framework that constructs shared intentionality prior to the evolution of language and that, altogether, these faculties constitute a "package" of capabilities. This claim is independent of the mechanism and biological implementation of that framework. Shared intentionality enables teamwork, which confers evolutionary advantage. Given shared intentionality, an individual can use it locally (i.e., communicate with herself) and that provides rationality; consciousness is awareness and control of these faculties in operation, and phenomenal consciousness arises so that sense experience can be communicated.

Second is the claim that shared intentionality requires mechanisms for abstraction, concretion, and explanation, and therefore Figure 3 is correct: there simply has to be a mechanism that abstracts part of the subconscious neural state into a succinct representation based on shared concepts and arranged as an explanation that can be communicated to others. An inverse mechanism, concretion, translates the explanation back into internal mental states that enrich the receiver's subconscious so that he can now deliver usefully cooperative behavior. The communicated explanation is not just a collection of concepts, it should employ structures to indicate logical operations ("and," "or"), temporal sequencing ("before," "then"), causation ("because," "in order to") etc., and the mechanisms for abstraction and concretion must be able to construct, deconstruct, and otherwise manipulate these. Explanations are wordless and communicated by mime, demonstration, or symbolic gestures and sounds, but I speculate they could provide a foundation for "mentalese" and, ultimately, language and speech.

Third is the claim that, to be effective, the explanation must be constructed using a model of the receiver's concretion operation: that is, a theory of mind with estimates of the receiver's knowledge and beliefs. To mechanize this, the sender can use suitable adjustments to her own concretion function, so that abstraction and generation of the explanation are achieved by an optimizing search to find an explanation whose adjusted concretion matches elements of the sender's own subconscious state. This argues that concretion and abstraction are closely associated and likely reside in the same mental unit, as portrayed in Figure 4. Predictive Processing is a biological process that can achieve these algorithmic requirements, whereby concretion and abstraction are represented by prediction and prediction error, respectively, and approximations to Variational Bayes use these to optimize abstract models and explanations. It is not necessary that these computational processes comprise or are precursors to a dual-process logical architecture, but it does seem plausible.

Fourth is the claim that local operation of the concretion/abstraction loop provides deliberation and rationality. Components of the loop have the ability to reason on explanations (as just described), and can recruit automated subconscious capabilities as subroutines via metaphorical translations (e.g., to reason about social hierarchy via translation to automated capabilities for understanding vertical space). Manipulation of the concretion operation (via its parameterization by a theory of mind) allows counterfactual and hypothetical reasoning. In their basic form, these capabilities "fall out" of the mechanisms for shared intentionality but may be enhanced with more sophisticated control, as portrayed in Figure 5.

Fifth is the speculative claim that awareness of attention to explanations and their associated models, together with control of their generation and interpretation, constitutes intentional consciousness. The claim explains the underlying purpose and construction of intentional consciousness but does not explain how or why it delivers a subjective experience: that aspect is speculative. I do not know of a way to confirm or refute this speculation but note that the underlying construction does explain otherwise puzzling facts, such as why the conscious mind is less an initiator of actions and more a reporter and interpreter of actions and decisions initiated elsewhere in the brain, why much of the experience of consciousness is of an inner dialog (rather than, say, a stream of images), and why the experience is at an intermediate level of representation. Thus, although the combination of Claims 1 to 5 explains the purpose of the Upper State and its models and the need for control and attention to its processes, I accept that they do not explain how these deliver consciousness as a nonphysical experience. (I do not apply to SIFT the illusionist step that some use to equate a report of consciousness with consciousness itself, although others are free to do so.)

Sixth is the claim that phenomenal consciousness is required so that we can communicate our sense experiences: "I just heard the roar of a lion." The content of some of our senses and subjective sensations (e.g., pains and emotions and moods) must be available for abstraction and incorporation into explanations for communication to others. Representations of these qualia are delivered to the abstraction/explanation mechanism where conscious resides, and therefore become conscious. The representations used for qualia are unimportant as long as they can be distinguished from one another, but we care about the representations actually used and refer to their experience as phenomenal consciousness.

These six claims are different from most other explanations and theories of consciousness and related topics in that they focus on functions needed to achieve certain purposes, and on mechanisms to deliver them. However, they are not incompatible with other theories and can be seen to provide context for several of them.

In particular, the combination of an abstraction function going "up" and a concretion function going "down," engaged together in an optimizing search to construct upper level models and communicable explanations, is consistent with predictive

processing models of brain function [28, 50, 69]. Thus, the biological plausibility of predictive processing and exploration of its neural correlates [70] can also support my claims (although these observations about PP are not unchallenged [126]).

Similarly, my presentation of proposed mechanisms derives a "dual process" architecture that is consistent with, and explains some aspects of, existing dual-process models of brain function [42, 48, 82]. Dual-process models hypothesize a logical, not physical, organization of the brain, and it is quite possible that they are realized by physical and neuronal mechanisms with quite a different structure, as hypothesized by global workspace theories [5, 32].

The Upper System in my dual-process model, performs calculations whose inputs and outputs are subconscious mental states: it is a part of the brain that senses and writes to other parts of the brain. This is consistent with Higher-Order Thought [17, 55, 124] and, when control of these processes is associated with attention, to Attention Schema Theory [60].

The six claims for SIFT are related and cumulative but differ in their scientific positioning. The fundamental criterion for a scientific theory is that it is testable and, moreover, falsifiable [115]. Schurger and Graziano make a further distinction: between scientific *laws* and *theories* [128]; both can be falsifiable but laws provide only descriptions (e.g., Newton's law of gravitation), whereas theories deliver explanations (e.g., General Relativity). They find that most theories of consciousness are, at best, laws and that only AST qualifies as a theory (although they concede it is a theory for *belief* in consciousness, rather than for consciousness itself).

Considering our six claims, the first (i.e., shared intentionality is the foundation for rationality and consciousness) is a testable and falsifiable theory about human evolution. The second and third (mechanisms for shared intentionality) are verifiable computer science constructions that deliver a theory about human cognition that seems testable by methods from psychology. The fourth (rationality) is a speculative theory, but it seems testable in both robots and humans. The fifth provides an explanation for the function and construction of the processes that underlie intentional consciousness. Their testability awaits development of more refined criteria for evaluating consciousness. However, this theory does not explain how or why intentional consciousness produces a subjective experience. The sixth claim is a logical consequence of the other five and delivers a strong explanation for the purpose and construction of phenomenal consciousness. Thus, SIFT fares at least as well as any description under Schurger and Graziano's criteria, and provides explanations that are both more comprehensive and more specific than other theories.

In conclusion, I propose that the Theory of "Shared Intentionality First" (SIFT) provides the most plausible basis for development of the cognitive package that includes consciousness and rationality. The theory explains the functions performed by intentional and phenomenal consciousness and how they are constructed. It reduces the "hard problem" for phenomenal consciousness to that for intentional

consciousness but, like all other theories of consciousness, it is unable to explain why the latter produces an apparently nonphysical experience. I invite others to develop or contest these proposals.

Acknowledgments and History

I am grateful to Antonio Chella and Owen Holland for advice and guidance to the literature. Owen Holland provided invaluable advice on previous work and relevant topics, and on framing these ideas for an audience beyond computer science. Harold Thimbleby alerted me to David Marr's work.

Colleagues at SRI, particularly Maria Paola Bonacina, Susmit Jha, Prashanth Mundkur, and N. Shankar have been patient and critical sounding boards. Prashanth provided extensive and helpful commentary. I greatly appreciate the support and encouragement of my boss, Pat Lincoln.

My subconscious first generated these ideas in July 2010 while reading Susan Blackmore's introduction [12]. I wrote them up in a note dated 24 July 2010 and gave an internal talk on the topic at SRI in May 2012. At that time, I mostly saw Figure 3 as an explanation for consciousness. As I studied related work I developed a more integrated treatment of consciousness and rationality and I wrote that up in May 2015 and gave a talk in June 2015, with revisions to the paper in October 2016. In Summer of 2017, I participated in a study on technology and consciousness and gave a talk in June 2017. Writing the workshop summary in 2018 [125] gave me the opportunity to reflect on many of these topics, and to understand the purpose of phenomenal consciousness. Believing that I now had a complete story, I revised the paper in early 2019. In what is probably a historic first, it was rejected by arXiv.org as being outside their scope. Rejection by a psychology journal educated me how (not) to communicate these ideas to psychologists. Susmit Jha and I presented a more technical paper based on these ideas at a AAAI workshop in 2019 [79]. I ruminated on these topics during Covid lockdown and the present paper is a fundamental rewrite developed over that period.

References

[1] Ralph Adolphs. The social brain: Neural basis of social knowledge. *Annual Review of Psychology*, 60:693–716, 2009. 9

[2] Igor L. Aleksander. Capturing consciousness in neural systems. In *International Conference on Artificial Neural Networks (ICANN 2)*, pages 17–22, Brighton, UK, 1992. 30

[3] Colin Allen and Michael Trestman. *Animal Consciousness*. In Max Vellmans and Susan Schneider, editors, *The Blackwell Companion to Consciousness*, chapter 5. Wiley, second edition, 2017. 24

[4] Simon D. Angus and Jonathan Newton. Emergence of shared intentionality is coupled to the advance of cumulative culture. *PLoS Computational Biology*, 11(10): e1004587, 2015. 22

[5] Bernard J. Baars. Global workspace theory of consciousness: Toward a cognitive neuroscience of human experience. *Progress in Brain Research*, 150:45–53, 2005. 3, 26, 35

[6] Simon Baron-Cohen. *The Evolution of a Theory of Mind*. In Michael Corballis C and Stephen E. G. Lea, editors, *The Descent of Mind: Psychological Perspectives on Hominid Evolution*. Oxford University Press, 1999. 9

[7] Roy F. Baumeister and E. J. Masicampo. Conscious thought is for facilitating social and cultural interactions: How mental simulations serve the animal-culture interface. *Psychological Review*, 117(3):945–971, 2010. 29

[8] Brent Berlin and Paul Kay. *Basic Color Terms: Their Universality and Evolution*. University of California Press, 1991. 21

[9] Derek Bickerton. *Language and Species*. University of Chicago Press, 1990. 4, 15

[10] Derek Bickerton. *Adam's Tongue: How Humans Made Language, How Language Made Humans*. Macmillan, 2009. 4, 28

[11] Jonathan Birch, Alexandra K. Schnell, and Nicola S. Clayton. Dimensions of animal consciousness. *Trends in Cognitive Sciences*, 24(10):789–801, 2020. 32

[12] Susan Blackmore. *Consciousness: A Very Short Introduction*. Oxford University Press, 2005. 36

[13] Piotr Boltuc. The philosophical issue in machine consciousness. *International Journal of Machine Consciousness*, 1(1):155–176, 2009. 31

[14] Charles J. Brainerd and Valerie F. Reyna. Gist is the grist: Fuzzy-trace theory and the new intuitionism. *Developmental Review*, 10(1):3–47, 1990. 19

[15] Michael E. Bratman. *Shared Agency: A Planning Theory of Acting Together*. Oxford University Press, 2013. 6, 28

[16] Joshua W. Brown. The tale of the neuroscientists and the computer: Why mechanistic theory matters. *Frontiers in neuroscience*, 8:349, 2014. 22

[17] Richard Brown, Hakwan Lau, and Joseph E. LeDoux. Understanding the higher-order approach to consciousness. *Trends in Cognitive Sciences*, 23(9):754–768, 2019. 3, 26, 27, 35

[18] Josep Call. Contrasting the social cognition of humans and nonhuman apes: The shared intentionality hypothesis. *Topics in Cognitive Science*, 1(2):368–379, April 2009. 22, 24

[19] Susan Carey. *The Origin of Concepts*. Oxford University Press, 2009. 23

[20] Sidney Carls-Diamante. The octopus and the unity of consciousness. *Biology & Philosophy*, 32(6):1269–1287, 2017. 25

[21] Sidney Carls-Diamante. Where is it like to be an octopus? *Frontiers in Systems Neuroscience*, 16, 2022. 25

[22] Peter Carruthers and Peter K. Smith, editors. *Theories of Theories of Mind*. Cambridge University Press, 1996. 9

[23] Olivia Carter et al. Conscious machines: Defining questions. *Science*, 359(6374):400–400, 2018. Letter in response to [33]. 31

[24] Charlotte Caucheteux, Alexandre Gramfort, and Jean-Rémi King. Evidence of a predictive coding hierarchy in the human brain listening to speech. *Nature Human Behavior*, 7:430–441, March 2023. 14

[25] David J. Chalmers. Facing up to the problem of consciousness. *Journal of Consciousness Studies*, 2(3):200–219, 1995. 20, 32

[26] Antonio Chella, Marcello Frixione, and Salvatore Gaglio. A cognitive architecture for robot self-consciousness. *Artificial Intelligence in Medicine*, 44(2):147–154, 2008. 30

[27] Antonio Chella and Riccardo Manzotti. Machine consciousness: A manifesto for robotics. *International Journal of Machine Consciousness*, 1(01):33–51, 2009. 30

[28] Andy Clark. Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(3):181–204, 2013. 3, 14, 23, 35

[29] Roger C. Conant and W. Ross Ashby. Every good regulator of a system must be a model of that system. *International Journal of Systems Science*, 1(2):89–97, 1970. 9, 23

[30] Leon de Bruin and John Michael. Prediction error minimization as a framework for social cognition research. *Erkenntnis*, pages 1–20, 2018. 29

[31] Leonardo de Moura and Nikolaj Bjørner. Satisfiability modulo theories: Introduction and applications. *Communications of the ACM*, 54(9):69–77, 2011. 16

[32] Stanislas Dehaene. *Consciousness and the Brain: Deciphering How the Brain Codes our Thoughts*. Penguin, 2014. 3, 26, 35

[33] Stanislas Dehaene, Hakwan Lau, and Sid Kouider. What is consciousness, and could machines have it? *Science*, 358(6362):486–492, 2017. 31, 38

[34] Daniel C. Dennett. *Consciousness Explained*. Penguin UK, 1993. 18

[35] Diane Derrien, Clémentine Garric, Claire Sergent, and Sylvie Chokron. The nature of blindsight: Implications for current theories of consciousness. *Neuroscience of Consciousness*, 1:1–14, 2022. 20

[36] Jean-Louis Dessalles. *Why We Talk: The Evolutionary Origins of Language*, volume 5. Oxford University Press, 2007. 28, 29

[37] Adrien Doerig, Aaron Schurger, and Michael H. Herzog. Hard criteria for empirical theories of consciousness. *Cognitive Neuroscience*, 12(2):41–62, 2021. 14

[38] Shona Duguid and Alicia P. Melis. How animals collaborate: Underlying proximate mechanisms. *Wiley Interdisciplinary Reviews: Cognitive Science*, 11(5), 2020. 24

[39] Robin I.M. Dunbar. The social brain hypothesis. *Evolutionary Anthropology*, 6(5):178–190, 1998. 28

[40] Ian J. H. Duncan. The changing concept of animal sentience. *Applied Animal Behaviour Science*, 100(1-2):11–19, 2006. 21

[41] Kresimir Durdevic and Josep Call. On the origins of mind: A comparative perspective. *Annual Review of Developmental Psychology*, 4:63–87, 2022. 24

[42] Jonathan St. B. T. Evans and Keith E. Stanovich. Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science*, 8(3):223–241, 2013. 15, 22, 26, 35

[43] Dean Falk. *Finding our Tongues: Mothers, Infants, and The Origins of Language*. Basic Books, 2009. 28

[44] Todd E. Feinberg and Jon M. Mallatt. *The Ancient Origins of Consciousness: How the Brain Created Experience*. MIT Press, 2016. 25, 26

[45] Frank A. Fishburn et al. Putting our heads together: Interpersonal neural synchronization as a biological mechanism for shared intentionality. *Social Cognitive and Affective Neuroscience*, 13(8):841–849, 2018. 22

[46] Charles W. Fox and Stephen J. Roberts. A tutorial on variational Bayesian inference. *Artificial Intelligence Review*, 38:85–95, 2012. 11

[47] Bruce A Francis and Walter M. Wonham. The internal model principle of control theory. *Automatica*, 12(5):457–465, 1976. 9

[48] Keith Frankish. Dual-process and dual-system theories of reasoning. *Philosophy Compass*, 5(10):914–926, 2010. 3, 15, 26, 35

[49] Keith Frankish. Illusionism as a theory of consciousness. *Journal of Consciousness Studies*, 23(11–12):11–39, 2016. 18

[50] Karl Friston. The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11(2):127, 2010. 13, 14, 35

[51] Karl Friston and Kiebel Stefan. Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society B: Biological sciences*, 364(1521):1211–1221, 2009. 13, 14

[52] Chris Frith. Consciousness is for other people. *Behavioral and Brain Sciences*, 18(4):682–683, 1995. 28

[53] David Gamez. Progress in machine consciousness. *Consciousness and cognition*, 17(3):887–910, 2008. 31

[54] Michael S. Gazzaniga. *Who's in Charge?: Free Will and the Science of the Brain*. Harper Collins, 2012. 19, 25

[55] Rocco J. Gennaro. *Higher-Order Theories of Consciousness: An Anthology*, volume 56 of *Advances in Consciousness Research*. John Benjamins Publishing, 2004. 3, 26, 27, 35, 45

[56] David Gil. Early human language was isolating-monocategorial-associational. In Angelo Cangelosi, Andrew D. M. Smith, and Kenny Smith, editors, *The Evolution of Language: Proceedings of the 6th International Conference (EVOLANG6)*, pages 91–98, World Scientific, Rome, Italy, March 2006. 15

[57] Peter Godfrey-Smith. *Other Minds: The Octopus, the Sea, and the Deep Origins of Consciousness*. Farrar, Straus and Giroux, 2016. 25

[58] Philip Goff, William Seager, and Sean Allen-Hermanson. *Panpsychism*. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2017 edition. 26

[59] Kirsty E. Graham, Claudia Wilke, Nicole J. Lahiff, and Katie E. Slocombe. Scratching beneath the surface: Intentionality in great ape signal production. *Philosophical Transactions of the Royal Society B*, 375(1789), 2020. 24

[60] Michael S. A. Graziano. *Consciousness and the Social Brain*. Oxford University Press, 2013. 14, 26, 27, 35

[61] Michael S. A. Graziano. Speculations on the evolution of awareness. *Journal of Cognitive Neuroscience*, 26(6):1300–1304, 2014. 27

[62] Michael S. A. Graziano. Attributing awareness to others: The attention schema theory and its relationship to behavioural prediction. *Journal of Consciousness Studies*, 26(3-4):17–37, 2019. 27

[63] Michael S. A. Graziano, Arvid Guterstam, Branden J. Bio, and Andrew I. Wilterson. Toward a standard model of consciousness: Reconciling the attention schema, global workspace, higher-order thought, and illusionist theories. *Cognitive Neuropsychology*, 37(3–4):155–172, 2020. 18, 26

[64] Richard L. Gregory. Perceptions as hypotheses. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, 290(1038):181–197, 1980. 14

[65] Peter G. Grossenbacher and Christopher T. Lovelace. Mechanisms of synesthesia: Cognitive and physiological constraints. *Trends in Cognitive Sciences*, 5(1):36–41, 2001. 21

[66] Jonathan Haidt. *The Righteous Mind: Why Good People Are Divided by Politics and Religion*. Vintage, 2013. Paperback edition. 3

[67] Stuart Hameroff and Roger Penrose. Consciousness in the universe: A review of the 'Orch OR' theory. *Physics of Life Reviews*, 11(1):39–78, 2014. 26

[68] Joseph Henrich. How culture made us uniquely human. *Zygon: Journal of Religion and Science*, 58(2):405–424, May 2023. 3

[69] Jakob Hohwy. *The Predictive Mind*. Oxford University Press, 2013. 3, 14, 35

[70] Jakob Hohwy and Anil Seth. Predictive processing as a systematic basis for identifying the neural correlates of consciousness. *Philosophy and the Mind Sciences*, 1(II), 2020. 14, 22, 35

[71] Owen Holland, editor. *Machine Consciousness*. Imprint Academic, 2003. 31

[72] Owen Holland. Forget the bat. *Journal of Artificial Intelligence and Consciousness*, 7(1):83–93, 2020. 32

[73] Owen Holland and Rod Goodman. Robots with internal models: A route to machine consciousness? *Journal of Consciousness Studies*, 10(4-5):77–109, 2003. 30

[74] David Hume. *A Treatise of Human Nature*. Three volumes, originally published by John Noon, 1739–40. 18

[75] Nicholas Humphrey. *Seeing Red: A Study in Consciousness*. The Belknap Press of Harvard University Press, 2006. 20, 24

[76] Nicholas Humphrey. *Sentience: The Invention of Consciousness*. Oxford University Press, 2022. 29

[77] Ray Jackendoff. *Consciousness and The Computational Mind*. The MIT Press, 1987. 21

[78] William James. *The Principles of Psychology*. Holt, New York, 1890. 13

[79] Susmit Jha and John Rushby. Inferring and conveying intentionality: Beyond numerical rewards to logical intentions. In Antonio Chella et al., editors, *Towards Conscious AI Systems Symposium (TOCAIS): AAAI Spring Symposium Series*, Stanford, CA, March 2019. Also available as arXiv:2207.05058. 26, 28, 36

[80] Susmit Jha, John Rushby, and N. Shankar. Model-centered assurance for autonomous systems. In António Casimiro et al., editors, *Computer Safety, Reliability, and Security (SAFECOMP 2020)*, Volume 12234 of Springer *Lecture Notes in Computer Science*, pages 228–243, Springer, Lisbon, Portugal, September 2020. 9, 11

[81] Nils B. Jostmann, Daniël Lakens, and Thomas W. Schubert. Weight as an embodiment of importance. *Psychological Science*, 20(9):1169–1174, 2009. 17

[82] Daniel Kahneman. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, 2011. 3, 15, 17, 26, 35

[83] Niki Kilbertus, Giambattista Parascandolo, and Bernhard Schölkopf. Generalization in anti-causal learning. arXiv:1812.00524, 2018. 11

[84] Robert Kirk. *Zombies*. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, spring 2019 edition. 32

[85] Christof Koch and Naotsugu Tsuchiya. Attention and consciousness: Two distinct brain processes. *Trends in cognitive sciences*, 11(1):16–22, 2007. 27

[86] Kurt Konolige et al. Centibots: Very large scale distributed robotic teams. In M.H. Ang and O. Khatib, editors, *Experimental Robotics IX: The 9th International Symposium on Experimental Robotics (ISER)*, number 21 in Springer Tracts in Advanced Robotics, pages 131–140, Singapore, 2006. 1

[87] Hermann Kopetz et al. *Emergence in Cyber-Physical Systems-of-Systems*. In Andrea Bondavalli, Sara Bouchenak, and Hermann Kopetz, editors, *Cyber-Physical Systems of Systems*, volume 10099 of *Lecture Notes in Computer Science*, pages 73–96. Springer-Verlag, 2016. 22

[88] George Lakoff. Mapping the brain's metaphor circuitry: Metaphorical thought in everyday reason. *Frontiers in Human Neuroscience*, 8:958, 2014. 17

[89] George Lakoff and Mark Johnson. *Metaphors We Live By*. University of Chicago press, 2008. First published 1980. 17

[90] Victor A.F. Lamme. Why visual attention and awareness are different. *Trends in Cognitive Sciences*, 7(1):12–18, 2003. 27

[91] Hakwan Lau. *In Consciousness we Trust: The Cognitive Neuroscience of Subjective Experience*. Oxford University Press, 2022. 27

[92] Benjamin Libet. Unconscious cerebral initiative and the role of conscious will in voluntary action. *Behavioral and Brain Sciences*, 8(4):529–539, 1985. 19, 24

[93] John Locke. *An Essay Concerning Human Understanding*. Edw. Mory, 1689. 27

[94] Francesco Marchi and Jakob Hohwy. The intermediate scope of consciousness in the predictive mind. *Erkenntnis*, 87(2):891–912, 2022. 21

[95] Hugo Gravato Marques, Rob Knight, Richard Newcombe, and Owen Holland. An anthropomimetic robot with imagination: One step closer to machine consciousness? In *Nokia Workshop on Machine Consciousness*, pages 34–35, 2008. 30

[96] Hugo Gravato Marques, Richard Newcombe, and Owen Holland. Controlling an anthropomimetic robot: A preliminary investigation. In *European Conference on Artificial Life*. pages 736–745, Springer, 2007. 30

[97] David Marr. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. W. H. Freeman and Company, 1982. 6, 21

[98] Massimo Marraffa. *Theory of Mind*. In James Fieser and Bradley Dowden, editors, *Internet Encyclopedia of Philosophy*. Unknown date. https://iep.utm.edu/theomind/. 9, 28

[99] Jennifer Mather. The case for octopus consciousness: Unity. *NeuroSci*, 2(4):405–415, 2021. Multidisciplinary Digital Publishing Institute. 25

[100] Jennifer A. Mather and Ludovic Dickel. Cephalopod complex cognition. *Current Opinion in Behavioral Sciences*, 16:131–137, 2017. 25

[101] Robin McKie. Humans hunted for meat 2 million years ago. https://www.theguardian.com/science/2012/sep/23/human-hunting-evolution-2million-years, 22 September 2012. 23

[102] Hugo Mercier and Dan Sperber. Why do humans reason? Arguments for an argumentative theory. *Behavioural and Brain Sciences*, 34(2):57–111, 2011. See also the commentary on page 74 by Roy F. Baumeister, E. J. Masicampo, and C. Nathan DeWall: "Arguing, Reasoning, and the Interpersonal (Cultural) Functions of Human Consciousness". 29

[103] Thomas Metzinger. *Being No One: The Self-Model Theory of Subjectivity*. MIT Press, 2004. 18

[104] Thomas Metzinger. Self models. *Scholarpedia*, 2(10):4174, 2007. 30

[105] Steven Mithen. *The Prehistory of the Mind: A Search for the Origins of Art, Science and Religion*. Thames and Hudson, London and New York, 1996. 24

[106] Henrike Moll, Ellyn Pueschel, Qianhui Ni, and Alexandra Little. Sharing experiences in infancy: From primary intersubjectivity to shared intentionality. *Frontiers in Psychology*, 12:667–679, 2021. 25

[107] Thomas Nagel. What is it like to be a bat? *The Philosophical Review*, 83(4):435–450, October 1974. 20

[108] Tihamér N. Nemes. *Cybernetic Machines*. Gordon and Breach, New York, 1970. English translation from Hungarian, originally published 1962. 30

[109] H. Penny Nii. The blackboard model of problem solving and the evolution of blackboard architectures. *AI Magazine*, 7(2):38–53, 1986. 13

[110] David A. Oakley and Peter W. Halligan. Chasing the rainbow: The non-conscious nature of being. *Frontiers in Psychology*, 8:1924, 2017. 29

[111] Cathal O'Madagain and Michael Tomasello. Shared intentionality, reason-giving and the evolution of human culture. *Philosophical Transactions of the Royal Society B*, 377(1843), 2022. 23

[112] Long Ouyang et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, Volume 35, pages 27730–27744, 2022. Also available as arXiv:2203.02155. 8

[113] Steven Pinker. *How the Mind Works*. Penguin UK, 2003. 17

[114] Arianna Pipitone, Francesco Lanza, Valeria Seidita, and Antonio Chella. Inner speech for a self-conscious robot. In Antonio Chella et al., editors, *Towards Conscious AI Systems Symposium (TOCAIS): AAAI Spring Symposium Series*, Stanford, CA, March 2019. 30

[115] Karl Popper. *The Logic of Scientific Discovery*. Routledge, 2014. First published in German 1934, English 1959. 35

[116] David Premack and Guy Woodruff. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4):515–526, 1978. 9

[117] Jesse Prinz. *The Intermediate Level Theory of Consciousness*. In Max Vellmans and Susan Schneider, editors, *The Blackwell Companion to Consciousness*, chapter 19. Wiley, second edition, 2017. 21

[118] Anand S. Rao and Michael P. Georgeff. *Modeling Rational Agents within a BDI-Architecture*. In Michael N. Huhns and Munindar P. Singh, editors, *Readings in Agents*, pages 317–328. Morgan Kaufmann, San Francisco, CA, 1997. 9

[119] Rajesh P. N. Rao and Dana H. Ballard. Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1):79–87, 1999. 14

[120] James A. Reggia. The rise of machine consciousness: Studying consciousness with computational models. *Neural Networks*, 44:112–131, 2013. 31

[121] Michael Rescoria. *The Language of Thought Hypothesis*. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2019 edition. 19

[122] David Robson. What is consciousness like for other animals and when did it evolve? *New Scientist*, July 7, 2021. 24, 25

[123] David M. Rosenthal. *Varieties of Higher-Order Theory*. In *Higher-Order Theories of Consciousness: An Anthology* [55], volume 56 of *Advances in Consciousness Research*, pages 17–44. 26

[124] David M. Rosenthal. *Consciousness and Mind*. Oxford University Press, 2005. 3, 27, 35

[125] John Rushby and Daniel Sanchez. Technology and consciousness. Technical report, Computer Science Laboratory, SRI International, Menlo Park, CA, September 2018. Minor update available as arXiv:2209.03956. 30, 36

[126] Tobias Schlicht and Krzysztof Dolega. You can't always get what you want: Predictive processing and consciousness. *Philosophy and the Mind Sciences*, 2, 2021. 35

[127] Britta Schünemann et al. Dogs distinguish human intentional and unintentional action. *Scientific Reports*, 11, 2021. 24

[128] Aaron Schurger and Michael Graziano. Consciousness explained or described? *Neuroscience of Consciousness*, 1:1–9, 2022. 35

[129] John R. Searle. *Collective Intentions and Actions*. In P. Cohen, J. Morgan, and M. E. Pollack, editors, *Intentions in Communication*, pages 401–416. MIT Press, 1990. 13

[130] John R. Searle. *Making the Social World: The Structure of Human Civilization*. Oxford University Press, 2010. 3

[131] Matthew Shardlow and Piotr Przybyła. Deanthropomorphising NLP: Can a language model be conscious? arXiv:2211.11483, November 2022. 31

[132] Aarohi Shrivastava et al. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. arXiv:2206.04615, June 2022. 31

[133] Arjen Stolk, Lennart Verhagen, and Ivan Toni. Conceptual alignment: How brains achieve mutual understanding. *Trends in Cognitive Sciences*, 20(3):180–191, 2016. 23

[134] Rebecca Wragg Sykes. *Kindred: Neanderthal Life, Love, Death and Art*. Bloomsbury Publishing, 2020. 24

[135] Joshua B. Tenenbaum, Charles Kemp, Thomas L. Griffiths, and Noah D. Goodman. How to grow a mind: Statistics, structure, and abstraction. *Science*, 331(6022):1279–1285, 2011. 8

[136] Michael Tomasello. *The cultural origins of human cognition*. Harvard University Press, 2009. 4, 28

[137] Michael Tomasello. *Origins of human communication*. MIT Press, 2010. 4, 28

[138] Michael Tomasello. *A Natural History of Human Thinking*. Harvard University Press, 2014. 23

[139] Michael Tomasello and Malinda Carpenter. Shared intentionality. *Developmental Science*, 10(1):121–125, 2007. 3, 6, 25, 28

[140] Giulio Tononi, Melanie Boly, Marcello Massimini, and Christof Koch. Integrated information theory: From consciousness to its physical substrate. *Nature Reviews Neuroscience*, 17(7):450–461, 2016. 26

[141] A. M. Turing. Computing machinery and intelligence. *Mind*, LIX(236):433–460, 1950. 31

[142] Jared Vasil, Paul B. Badcock, Axel Constant, Karl Friston, and Maxwell J.D. Ramstead. A world unto itself: Human communication as active inference. *Frontiers in Psychology*, 11:417, 2020. 10, 23, 28

[143] Hermann von Helmholtz. *Handbuch der Physiologischen Optik III*, volume 9. Verlag von Leopold Voss, Leipzig, Germany, 1867. 14

[144] Ashley Ward. *Sensational: A New Story of Our Senses*. Profile Books, London, 2023. 20

[145] Wanja Wiese and Thomas K. Metzinger. *Vanilla PP for Philosophers: A Primer on Predictive Processing*. In Thomas K. Metzinger and Wanja Wiese, editors, *Philosophy and Predictive Processing*, chapter 1. MIND Group, Frankfurt am Main, 2017. 11, 14

[146] *Alice and Bob*. Wikipedia. https://en.wikipedia.org/wiki/Alice_and_Bob. 5

[147] *The Dress*. Wikipedia. https://en.wikipedia.org/wiki/The_dress. 21

[148] *Reflective Programming*. Wikipedia. https://en.wikipedia.org/wiki/Reflective_programming. 27, 30

[149] Alan F. T. Winfield. Experiments in artificial theory of mind: From safety to story-telling. *Frontiers in Robotics and AI*, 5, June 2018. Article 75. 30

[150] Thomas R. Zentall et al. Concept learning in animals. *Comparative Cognition & Behavior Reviews*, 3:13–45, 2008. 23