# Indiana Undergraduate Journal of Cognitive Science

An Online Journal of Research and Writing in Cognitive Science

---

## Contents

### Articles

<table>
<thead>
<tr>
<th>Article</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Information, Relevance, and Objectivity</td>
<td>3</td>
</tr>
<tr>
<td><em>Logan Fletcher, University of Toronto</em></td>
<td></td>
</tr>
<tr>
<td>Consciousness (Response to the Hard Problem)</td>
<td>23</td>
</tr>
<tr>
<td><em>Artem Tsvetkov, Carleton University</em></td>
<td></td>
</tr>
<tr>
<td>The Joining of Neuroscience, Psychology, and Philosophy in a Search for the Self</td>
<td>28</td>
</tr>
<tr>
<td><em>Samantha Noll, West Chester University of Pennsylvania</em></td>
<td></td>
</tr>
<tr>
<td>Searle's Chinese Room Argument and its Replies: A Constructive Re-Warming and the Future of Artificial Intelligence</td>
<td>51</td>
</tr>
<tr>
<td><em>Mariano de Dompablo Cordio, University of Wisconsin, La Crosse</em></td>
<td></td>
</tr>
<tr>
<td>Thinking About Your Foolishness: Foolishness in Terms of Metacognition</td>
<td>58</td>
</tr>
<tr>
<td><em>Jelena Markovic, University of Toronto</em></td>
<td></td>
</tr>
</tbody>
</table>

### General Information

<table>
<thead>
<tr>
<th>Information</th>
<th>Page</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Information</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Author / Submission Instructions</td>
<td>67</td>
</tr>
</tbody>
</table>

**On the Cover:** Neuron © 2007 – Jason Lam. Art Used with Permission.
Fred Dretske opens his book *Knowledge and the Flow of Information* with the assertion that information is *objective*. In contrast to the common view of information “as something that depends on the interpretive efforts—and, hence, on the prior existence—of intelligent life,” Dretske intends to develop a view of information “as an objective commodity, something whose generation, transmission, and reception do not require or in any way presuppose interpretive processes,” something “independent of its actual or potential use by some interpreter” (Dretske, 1981, p. vii).¹ Dretske has a significant stake in the question of information’s objectivity, because he is trying to use information to naturalize the mind:

> Can you bake a mental cake using only physical yeast and flour? The argument is that you can. Given the sort of information described… something the most reflective materialists should be willing to give, we have all the ingredients necessary for understanding the nature and function of our cognitive attitudes. (1981, p. xi)

And, as Dretske points out elsewhere, “One cannot have a recipe for cake… that lists a cake as an ingredient…. This is why a recipe for thought cannot have interpretive attitudes or explanatory stances among the eligible ingredients” (2000, p. 209). In other words, if information is *not* objective, if it does depend crucially on interpretation, this effectively renders Dretske’s whole naturalization project viciously circular. As we might expect, then, Dretske takes up the challenge, in the remainder of the book and in other writings, of defending his claim that information is an “objective commodity”, by arguing that everything else on which information depends is itself objective.

---

¹ Throughout, ‘objective’ will be taken in Dretske’s sense to mean *interpretation-independent*, and will take on a more specific meaning as we progress. There may of course be other sorts of objectivity that *are* interpretation-dependent, but these will not be discussed.
It is surprising, then, to say the least, when in chapter 5, Dretske admits that whether or not a signal carries a piece of information is “a question that may not have an objectively correct answer… an issue that is, in part at least, responsive to the interests, purposes, and, yes, values of those with a stake in the communication process” (1981, p. 132-3). This would seem, on the face of things, to be utterly inconsistent with Dretske’s earlier claim that information is objective. Information, one might think, either is or is not objective. So what is going on here? Has this inconsistency simply gone unnoticed by Dretske? It has not; Dretske has a rather subtle story to tell about how these apparently opposing positions can be reconciled, and how we can consistently hold that information is indeed objective in the sense required for its use in a non-circular naturalization of the mind, while admitting that information is, in a distinct sense, relative to subjective interpretation. It is the topic of the present paper to examine this story, to see whether it holds up to close scrutiny. I will argue that it does not: The required separation between the subjective and objective aspects of information cannot be maintained; in the case of information, we cannot sequester the subjective factors in such a way as to maintain a space of objectivity purified of subjective influence. Subjective interpretation is inextricably entangled with every aspect of the informational phenomenon, leaving us no way to ‘factor out’ the subjectivity by relativizing information to an explicit standard. This has the result, I will argue, that Dretske’s project of using information to naturalize the mind is unavoidably circular.

**Absolute Information**

The idea of information, as understood by Dretske, is fairly easy to grasp. Dretskean information is roughly equivalent to Grice’s (1957) natural meaning, the sense in which smoke means fire. In Dretske’s terms, smoke carries the information that there is fire. Or, to take one of Dretske’s own examples, a pressure gauge carries information about the pressure within a certain boiler (1981, p. 123). In general, a signal (e.g. the pressure gauge being such-and-such) carries the information that conditions at the source are so-and-so (e.g. that the pressure is so-and-so) if and only if it is certain, given the signal being such-and-such, that the conditions at the source are indeed so-and-so. Here is how Dretske explicitly defines the informational content carried by a signal about a source:

A signal \( r \) carries the information that \( s \) is \( F \) = The conditional probability of \( s \)’s being \( F \), given \( r \) (and \( k \)), is 1 (but, given \( k \) alone, less than 1) (1981, p. 65)
Loewer points out that Dretske’s invocation of probabilities is problematic, because there is no available account of probability that will meet Dretske’s needs, and Dretske himself fails to provide one (Loewer, 1983). As Loewer acknowledges, however, the appeal to probability per se is not crucial to Dretske’s account, and as Dretske himself explains, “A conditional probability of 1 between $r$ and $s$ is a way of describing a lawful (exceptionless) dependence between events of this sort,” which “nomically precludes $r$’s occurrence when $s$ is not $F$” (1981, p. 245). Dretske only ever distinguishes two “probabilities”, namely $p=1$ and $0<p<1$, which, as far as I can tell, might just as well be called “necessity” and “possibility”. What is important, then, is the nomic regularity that relates $r$ to $s$, which need not be explained in terms of probabilities.

A more serious concern is the presence of “$k$” in the definition, which represents “what the receiver already knows… about the possibilities that exist at the source” (Dretske, 1981, p. 65). The motivation for this is the idea that if I already know that certain possibilities at the source are excluded, and you do not, I may receive a piece of information from the signal that you fail to receive from it. Dretske considers a shell game in which a peanut is under one of four shells: If I know the peanut is not under shells 1 and 2 (I have seen they are empty), and you do not know this, then, when we both see that shell 3 is empty, I receive the information that the peanut is under shell 4, whereas you do not (1981, p. 78-9). The definition of informational content in terms of knowledge would seem to compromise the supposed objectivity of information; in fact, it appears to lead straightforwardly to circularity, since Dretske goes on to analyze knowledge as information-caused belief (1981, p. 86). Dretske anticipates this potential criticism, arguing that the apparent circularity is in fact a finite regress, in which we can recursively analyze any presupposed knowledge into its informational basis, until we eventually “reach the point where the information received does not depend on any prior knowledge about the source” (1981, p. 86-7). At the point where this recursive process bottoms out, we will be left with what Steven Savitt calls “absolute informational content”, which is defined just as in Dretske’s original definition, but without the parenthetical reference to $k$ (Savitt, 1987). Cohen and Meskin (2006) express doubts about whether this recursive process will always bottom out into absolute information in the way Dretske suggests, but it is worth noting that there is a much more direct way to reduce knowledge-dependent information to absolute information, which is simply to consider the absolute correlate of the knowledge-relative informational content: the informational content carried by the same signal without taking $k$ into
account. To be sure, this is not a proper ‘reduction’, since the absolute correlate will generally be less specific in its content than the knowledge-dependent information, but, as Savitt notes, we can still explain cases such as the discrepancy between observers in the shell game, by positing an inferential process by which we derive the knowledge-dependent content (that the peanut is under shell 4) from the absolute content (that shell 3 is empty) plus our prior knowledge (that shells 1 and 2 are empty). There are, then, good reasons to doubt that the inclusion of k in Dretske’s definition seriously compromises the objectivity of information, so defined. But in any case, Dretske’s theory includes a theory of absolute information. Therefore, I will set aside the question of k-relativity for the rest of the paper, restricting my attention exclusively to absolute information. Should this fail to be objective, it will create much more serious problems for Dretske’s project.

Channel Conditions

As Dretske’s definition makes clear, for a signal to carry information about a source just is for the appropriate nomic regularity to hold between them: Given that the signal is such-and-such, necessarily the source is so-and-so. Crucially, this nomic regularity supports the following counterfactual claim: Were it not the case that the source is so-and-so, it would not be the case that the signal is such-and-such. It is this lawful quality—indicated by ability to support counterfactuals—that distinguishes informational relationships from mere de facto correlations, a distinction stressed by Dretske as crucial (1981, p. 73-5). Dretske at one point glosses these nomic regularities as “natural laws” (1981, p. 77-8), but they are strikingly different from familiar laws of nature. Laws of physics, for instance, are thought to express necessary relationships because they have never been observed to admit of exceptions. In contrast, it is strange to think that there is something necessary about the relationship between the pressure gauge and the boiler pressure. We do our best to maintain this relationship, of course, but it is not difficult to imagine cases in which it fails to hold, for instance when the gauge is damaged, or disrupted by a strong magnetic field, or simply disconnected—that these possibilities are not the case is of course not something we can tell simply by reading the gauge itself. It would therefore seem to be a contingent matter whether or not the signal tracks conditions at the source, not any sort of law.

Dretske explains that the informational relationship between signal and source holds only relative to a set of channel conditions (1981, p. 115). In the example of the pressure gauge, the set of channel conditions will consist of all the facts, other than the fact that the boiler pressure is so-and-
so, on which the such-and-such-ness of the pressure gauge depends: That the gauge is hooked up properly to the boiler, that there are no strong magnetic fields interfering, and so forth. It would seem that the complete set of channel conditions would have to be infinite, since no matter how long our list grows, there will always be some scenario, however far-fetched, in which the signal is such-and-such, all the channel conditions already listed are met, and still the source is not so-and-so. Now, one might expect Dretske to say that if each of these facts is true (that the gauge is hooked up properly, that there are no strong magnetic fields, etc.), and the signal is such-and-such, then the signal will carry the information that the source is so-and-so. Dretske does not say this, and for good reason. The problem is not that the source will not be so-and-so (it will), but that, in the envisioned scenario, there is no way to distinguish the source from the channel conditions. Since changing any one of these conditions will affect the signal, there is no objective fact about which one is actually the source—we can treat any of them as the source, depending on our interests.\(^2\) Clearly, this sort of relativism will be unacceptable to Dretske.

Here is what Dretske does say: The channel conditions (for absolute information) are the “existing conditions (on which the signal depends) that… generate no (relevant) information” (1981, p. 115). What does it mean for a condition to “generate no (relevant) information”? It means, first, that the fact associated with the condition (“there are no magnetic fields interfering”) is actually true, and, second, more problematically, that there are no relevant alternative possibilities to it being true. For a channel condition to hold, then, requires more than just something being true of the actual world—there is an additional modal requirement. Hence it not only must be true that there is no magnetic field interfering, but also the possibility of a magnetic field interfering must not be relevant. It is by appealing to this modal aspect that Dretske is able to distinguish between the source and the channel conditions; the source has relevant alternative possibilities (which nomically covary with conditions at the signal), whereas the channel conditions do not.

What does it mean for a possibility to be relevant or irrelevant? If there is a magnetic field that unpredictably turns ON and OFF, interfering with the gauge’s readings only when it is ON, then Dretske would surely say that the gauge will not ever carry information about the boiler pressure, even

\(^2\) We cannot, however, treat all of them as the source at the same time, since there must be a channel in order for there to be a source (Dretske, 1981, p. 117). It makes no sense for the signal to carry information about all the conditions on which it depends.
when the magnetic field is OFF, because it will always be a relevant possibility that it is ON. On the other hand, if a magnetic field has never interfered with the boiler gauge, then Dretske would probably consider the possibility of magnetic interference to be an irrelevant one. As Dretske acknowledges (1981, p. 130), however, de facto frequencies of this sort, covering actual past occurrences, are not criterial of whether or not something is a relevant possibility. Indeed, they cannot be criterial, since this would undermine Dretske’s central claim that informational relationships are nomic regularities, as opposed to de facto correlations. So what does make something a “relevant possibility”?

Relational Objectivity

Whether or not the signal carries information about the source depends on whether or not the channel conditions hold. Hence, it must be an objective truth that the conditions hold, if the information-carrying relationship they support is itself to be objective. But as we’ve seen, Dretske is forced to include a modal element in the definition of channel conditions: Whether or not a channel condition holds is in turn dependent upon what relevant alternative possibilities are present in the actual situation. Does it make sense for a possibility to be actually present, and for this to be an objective fact? Dretske thinks it does:

The difference between a relevant and an irrelevant alternative resides, not in what we happen to regard as a real possibility (whether reasonably or not), but in the kind of possibilities that actually exist in the objective situation. (Dretske, 1981a, p. 377)

Dretske discusses some examples that he takes to illustrate the actual presence and absence of relevant possibilities. For instance, Dretske considers a bird-watcher who identifies a bird as a Gadwell duck. Supposing that the bird actually is a Gadwell, does the bird-watcher know that (i.e. possess the information that) it is a Gadwell? Is the possibility that it is a look-alike Siberian Grebe a relevant alternative? (1981a, p. 369) It depends, Dretske argues, on the actual facts: The possibility is surely relevant if there are in fact Grebes in the vicinity. On the other hand, it is irrelevant, according to Dretske, if Grebes do not actually exist, or if “due to certain geographical barriers, they are confined to their Siberian habitat” (1981a, p. 377). Then there are borderline cases, which we can produce ad

---

3 This is, of course, a question about k-relative, not absolute, information. This is incidental, however—the point being made concerns relevant alternative possibilities in general.
nauseum: What if the Grebes remain in Siberia, but there is no barrier preventing them from migrating? (1981a, p. 377) What if a hunting organization has plans to populate the area with Grebes, but has not yet been given governmental permission to do so? How serious must these plans be, to give rise to a relevant alternative? Suppose they have already moved caged Grebes into the area—do we trust them to keep the birds locked away safely? What if one of the handlers is known to be careless? And what sort of carelessness, exactly, is appropriate to generalize to apply to the specific case of Grebe-handling?

It is cases such as these, apparently unbounded in their potential for subtle contextual entanglements, that impel Dretske to acknowledge a pragmatic element to the presence of the information-carrying relationship. Dretske admits that such cases make information-carrying “a question of degree” (1981, p. 132), but it is difficult not to feel that in admitting this, he is admitting too little. Dretske describes “the difficult question” as “the question of when an alternative… is just too remote to qualify as relevant” (1981a, p. 376), but it would seem that the more challenging pragmatic problem would be to determine the metric of remoteness that this formulation presupposes. Is a Grebe migration from Siberia more remote a possibility than a hunting organization realizing their half-hearted plan to import Grebes? Whichever way your intuitions may tend on this matter, it is undoubtedly a question whose answering requires the balancing of two content-rich complexes of reliability-cum-variability that are not commensurable with each other in any straightforward way. In contrast, the pragmatic problem discussed by Dretske—that of determining what degree of remoteness is sufficient to render a possibility irrelevant—would seem to be precisely analogous to the problem of determining how tall a person must be in order to be considered a “tall” person, a question that may well involve some interesting sensitivity to context (relativizing to age, gender, or race, for instance), but which clearly lacks the rich entanglements of questions regarding the relative remoteness of possibilities.

Bearing this in mind, here is how Dretske resolves the issue of the acknowledged pragmatic element to relevant possibilities: He claims that being a relevant alternative possibility (hence also being an information-carrying relationship) is an absolute matter, determined relative to a pragmatically-constituted standard of what counts as relevant. The case is analogous, Dretske argues, to the question of whether something is flat: Although ‘flat’ is an “absolute concept”, meaning that “nothing can be flat if it has any bumps and irregularities,” flatness is determined relative to “what counts as a bump or irregularity” (1981a, p. 366). Dretske calls concepts of
this sort “relationally absolute”, and claims that relevant possibilities and information-carrying relationships have a relationally absolute character (1981a, p. 367). Information-supporting nomic regularities, then, are indeed exceptionless, permitting no relevant possibilities, but they must be understood relative to a standard of relevance.

So much for the absoluteness of information; what about its objectivity? Is it relational in the same manner? Dretske’s position here is not so easy to discern. On the one hand, when faced with the difficulties of determining the relevant possibilities in challenging borderline cases, Dretske admits that the “question may not have an objectively correct answer” (1981, p. 132-3), thereby suggesting that what objectivity it does have will need to be understood relationally. And when Dretske speaks of seeking relevant alternatives in the “possibilities that actually exist in the objective situation” (1981a, p. 377), the “objective situation” is clearly one constituted by a standard of relevance. Otherwise, relevant possibilities would exist independent of such a standard, and then what would we need the standard for? On the other hand, Dretske states that a skeptic, “someone with very high standards, someone who considers almost any alternative relevant,” would “speak falsely” by making claims about (the failure of) information transmission on the basis of such high standards (1991, p. 192). But if relevant possibilities are determined relative to a standard of relevance, then on what basis does Dretske assert that the skeptic’s standard is mistaken, if not by appealing to a second-order standard that renders the first-order standard redundant? Dretske’s answer is that when one makes claims about the success or failure of information transmission, it is the “background of standard or intended use [of information] to which the claim is relative… A speaker who brings a novel or unusual viewpoint to the situation being described… is expected to indicate that difference in order to avoid misunderstanding” (1991, p. 194). Dretske, then, is not providing an argument that the skeptic is objectively wrong to choose an eccentric standard of relevance; such an argument would be absurd, since the appeal to “standard or intended use” as an objective means of disqualifying the skeptic’s standard would directly contradict Dretske’s description of information, qua objective, as “independent of its actual or potential use by some interpreter” (1981, p. vii). Rather, Dretske is making the thoroughly

---

4 David Lewis makes the same point in a very similar context, arguing that a man who “attached eccentric relative importances to respects of comparison of worlds” would not be entitled to make claims on the basis of this standard, “at least not without giving warning of his eccentric notions.” Lewis claims the eccentric individual “lies”, because “he temporarily changes the conventional meaning of his words” (Lewis, 1979, p. 93-4).
pragmatic claim that the skeptic’s standard lies outside the range of conventional standards, which are more reasonable for identifying information as it is conventionally used.

The objectivity of information, then, is indeed to be understood relationally, as relative to a standard of relevance, just as the absoluteness of information is so relative. The skeptic’s determinations of relevant possibilities, then, will be objective relative to the skeptic’s standard, just as any other set of determinations will be objective relative to its own standard. This means that the standard favored by Dretske will necessarily be arbitrary from the perspective of objectivity itself—this is an unavoidable consequence of understanding relevant possibilities in relation to a standard of what counts as relevant. Nevertheless, Dretske can still argue, and I take him to be arguing, that the standard that is pragmatically reasonable (or, perhaps, any standard within the set of pragmatically reasonable standards) provides a basis for objective claims about the transmission of information. Just because the standard of relevance that effectively defines information transmission is chosen on the basis of pragmatic issues regarding the use of information, does not imply that information, so defined, is not objective. Were the definition itself to make essential reference to information use, objectivity would surely be compromised, but this is a completely different matter from information use playing an essential role in motivating the choice of definition. The difference is that between defining length in terms of “reasonable units” and defining length in terms of “metres” or “feet”—the latter standards, unlike the former one, provide a foundation for objective claims about length, in spite of the subjective and pragmatic concerns that may have figured in the decision to adopt them in the first place.

As suggested by the example of length, a paradigmatically objective property, all objective claims will at some level exhibit this standard-relativity. For a claim to be objective is just for there to be a clean interface—an unambiguous standard—joining together the subjective matter of what counts for deciding the claim, and the objective matter of how the claim is decided, given what counts, in a manner that enforces their mutual separation. Subjectivity comes into play in the formulation of a standard, but the application of a standard, once formulated, needs no further help from subjective interpretation—it applies, so to speak, all on its own. For example, in playing a game, we can have diverging subjective opinions regarding what should count as a legal or illegal move.5 Such subjective

5 Conossieurs of spontaneous game invention will be familiar with cases in which the agreed-upon rules of a newly invented game turn out not to anticipate every situation that
matters, however, are (in most games, at least) strictly limited to issues concerning the formulation of the rules, which, once formulated, provide a clear standard that objectively determines issues of legality. There is then a sharp distinction between the subjective, value-laden, and, I am tempted to say, aesthetic issues involved in determining what game we are playing, and the purely objective issues about what is and is not the case with respect to legality, given that we are playing this game. The rules, once formulated, can be applied objectively, that is to say, without any concern for the subjective issues that figure in why the rules are as they are. There is a clear line where subjectivity leaves off, and objectivity takes over.

This, I think, is how Dretske wants us to understand the nature of information as simultaneously pragmatic and objective. Dretske wants to say that subjective issues determine the standard by which possibilities are judged to be relevant (what game we are playing), and that the facts about facts about which ones are relevant (which moves are legal), given this standard (given that we are playing this game), are entirely objective, leaving no leeway for subjective interpretation. But we might reasonably worry that this sort of “objectivity” is vacuous, since we can always posit a standard according to which our claims are “objective”, regardless of how paradigmatically subjective these claims are. Consider, for instance, if Dretske had argued that beauty is objective, relative to a standard of what counts as beautiful. Our own interests and values, he says, figure in the determination of our personal standard of beauty, but given this standard, there are perfectly objective facts about what is and is not beautiful. I take it that Dretske does not want to argue that information is objective in the same sense that beauty is, but what prevents us, then, from saying that beauty is relationally objective? The answer, I think, is that there is no such thing as a relationally objective standard of beauty, not even of Dretske-beauty, beauty as judged by Dretske. Beauty is so richly intertwined with subjectivity, so unavoidably tacit, that it is just not possible to ‘factor out’ the subjective element by formulating a standard that can be applied in abstraction from this subjective vantage point. Anything that could genuinely be said to distinguish beautiful things would have to appreciate this beauty, and hence would need to be, in some sense, a subject, with an ineliminably implicit grasp of aesthetic value. Or so, at any rate, I would claim.

Whether or not you agree with me about beauty, I think it is clear that if the evaluation of a claim is indeed dependent on such ineliminably

emerges within play. There may in such cases be opposing sets of strong feelings regarding which potential extension of the rules is most “true to the spirit” of the game.
implicit elements, the claim cannot be said to be an objective one. This consideration leads directly to what we might call the Explicitness Condition: In order for a standard to serve as a basis for making (relationally) objective claims, that standard must be able to be formulated in an explicit manner. There must be no implicit factors determining how the standard is to be applied; such implicitness must be cashed out, by rendering these factors explicitly, as part of the formulation of the standard. If the application of the standard depends on factors that are systematically resistant to explicit formulation, then this suggests that the standard fails as a basis for making objective claims. The idea behind the Explicitness Condition is that if a standard is to support relational objectivity, its application can neither depend upon nor leave room for subjective interpretation. If it is in fact interpretation-independent, it ought to be amenable to explicit expression.6

The question of the objectivity of Dretskean information, then, comes down to the question of whether the conventional or pragmatically reasonable intuitions about relevant alternative possibilities can be captured in a standard that meets the Explicitness Condition. Dretske himself does not attempt to articulate anything approaching such a standard, and no available theories of relevant alternatives seem to make any progress towards discharging the implicit and pragmatic aspects. David Lewis’s theory of relevant alternatives (Lewis, 1996), for instance, described as “the most sophisticated version… to date” (Shaffer, 2001, p. 202), makes essential reference to “salient resemblance” in analyzing relevant alternatives, which does not leave us any better off with respect to the task of eliminating the dependence on pragmatic elements. It is worth pointing out that for most theorists, Dretske included, the primary use to which relevant alternatives are put is in arguments against the skeptical claim that knowledge is nonexistent or severely limited. As Stewart Cohen points out,

6 There will of course be skeptical worries about whether any standard could ever be formulated in an entirely explicit manner. The skeptic will claim that there will always be factors on which the application of the standard depends, that remain necessarily implicit. To some extent, this concern will be addressed by the discussion of the mechanization of the application of standards in the last section of this paper. For now, I will just point out that we can, if we wish, interpret the Explicitness Condition as a regulative ideal, so that the Condition just requires that the standard can be rendered arbitrarily explicit, depending on how carefully and conscientiously we proceed in its formulation. On this interpretation, even if any actual formulation of a standard will always depend on residual implicit elements, the standard can still meet the Explicitness Condition.
the notion of a relevant alternative can do this job perfectly well, even if we are unable “to provide a precise formulation of the criteria of relevance” (1991, p. 33), because here there is no prior commitment to the interpretation-independent character of knowledge (or information). If, on the other hand, “we want the theory of relevant alternatives to provide an analysis of knowledge”—as Dretske of course does, via the notion of information—“then the failure to provide precise criteria would constitute a failure of the theory” (Cohen, 1991, p. 33).

Although it seems to me that the burden lies with Dretske to show that we can formulate an explicit standard for determining relevant alternatives, the lack of any extant accounts that accomplish this surely does not demonstrate that we cannot. There are, however, reasons to doubt that we can formulate such a standard, reasons to think that the presence of relevant alternatives, like the presence of beauty, is an inexorably subjective matter. For one thing, in spite of Dretske’s insistence that the issue of which possibilities are relevant is “a question of degree”, there is not any straightforward manner of assessing the degree of remoteness of possibilities. As we observed earlier, relative remoteness cannot be assessed without becoming embroiled in complex issues regarding the forms of reliability and variability that are present in situations. Dretske tells us that

To qualify as a relevant possibility, one that actually affects the equivocation of (and therefore the information in) a signal, the possibility envisaged must actually be realizable in the nuts and bolts of the particular system in question. (1981, p. 131).

The suggestive metaphor of “the nuts and bolts of the system” brings to mind the picture of certain basic pieces that, by virtue of their shapes, place constraints on the manner with which they can be combined. The problem is that in the real-world situations in which Dretske wants information to be carried, it is not obvious which is the relevant decomposition into nuts and bolts. What counts as a “bolt”, or as a “hinge point” is itself extremely sensitive to context. Recall the question about what sort of carelessness is appropriate to generalize to the case of Grebe-handling. It is difficult to see how such a question could be answered without delving into issues about the nature of carelessness. But it is also not clear that the sort of carelessness that is relevant to Grebe-handling is a detachable “piece” that would arise in other contexts as well. The nearness of the possibility that the Grebe-handler will fail to properly secure the Grebe’s cage may well depend on an idiosyncratic constellation of factors, including, for instance, competence with locks and cages. But what sort of cage-competence generalizes to the current situation? For any particular situation, we will be able to ask
indefinitely many such questions, regarding the \textit{proper fit} of regularities to the concrete situation. The standard of relevance, though, if it is to serve as a basis for talking about “information” \textit{in general}, must provide an answer, not only to all such questions in any particular situation, but to all such questions in \textit{all} of the situations in which we want to talk about information. With similar considerations in mind, Cohen says that “Because there are so many complex and controversial variations in examples like this, it is exceedingly difficult to capture the distinction between relevant and irrelevant alternatives in a precise criterion” (Cohen, 1991, p. 32). Indeed, the task of formulating an explicit standard of relevance seems prohibitively difficult.

\textbf{Modal Complexity}

At this point, I want to suggest that the difficulties we encounter in evaluating claims about relevant possibilities can fruitfully be viewed as endemic to a larger class of claims, what I will call \textit{modally complex} claims, of which relevant possibility claims are a subset. Setting aside all claims of an \textit{obviously} normative or evaluative nature, such as moral or aesthetic claims, I propose that we can divide much of what remains into three classes. First, there are \textit{non-modal} claims: These are claims about the \textit{actual} world, stripped of all modal character, that is, independent of any facts about what \textit{could be}, \textit{would have been}, or \textit{almost is} the case. We do not need to appeal to possible worlds at all in assessing non-modal claims. Claims about the mass of something, or the relative positions of two things, are paradigmatic examples. Second, there are \textit{modally simple} claims, claims about what is \textit{possible} or \textit{necessary}, either absolutely or conditionally. Modally simple claims make no reference to the actual world, and hence can be assessed without considering what is \textit{actually} the case. Absolute possibility, under most interpretations of modality, would be identified with \textit{logical} possibility, since logical truths are usually thought to hold in all possible worlds. Claims about conditional possibility, for instance claims about what is \textit{physically} possible, are directed to certain subsets of the set of possible worlds. Third, there are \textit{modally complex} claims: These are claims directed to what we might think of as the \textit{modally-animated actual world}, that is, the actual world, considered as situated amongst surrounding possibilities. Counterfactual conditionals and claims about the nearness and remoteness of possibilities are modally complex—they are true or false at
the actual world, but they can only be assessed by considering the relation of the actual world to other possible worlds.\footnote{Strictly speaking, modal complexity does not require any reference to the actual world, since we can treat other possible worlds as modally animated, and evaluate counterfactuals, for instance, at these worlds. But in the vast majority of cases of concern, the world at which modally complex claims are true or false is the actual world.}

Non-modal claims and modally simple claims are both potentially capable of relational objectivity, that is, we can formulate standards, on the basis of which we can understand such claims, which meet the Explicitness Condition. For instance, claims about the lengths of various things can be objective relative to a standard that expresses how the terms by which we measure lengths are to be applied. Or claims about what is possible given, say, Newtonian physics can be understood relative to an explicit statement of the appropriate laws and fundamental entities (here we actually specify what the “nuts and bolts” are). In fact, the categories of non-modal and modally simple claims appear to exhaust the claims to which empirical science and logic, respectively, are committed.

The standards for evaluating modally complex claims, on the other hand, are systematically resistant to explicit formulation. This is a consequence of the fact that modally complex claims depend crucially on relations of similarity between the actual world and other possible worlds, which relations may shift dramatically depending on the particular claim at issue. David Lewis defends the view that the “relative importances of respects of comparison,” and hence also the relations of “comparative overall similarity” between worlds, while volatile, are “at least roughly fixed,” and can be treated as such for the purposes of evaluating modally complex claims (Lewis, 1973, p. 92-3). But consider the counterfactual conditional with which Lewis opens his book: “If kangaroos had no tails, they would topple over” (1973, p. 1). This claim reveals the basic nature of counterfactuals: A change is made to the actual world, which will have further consequences, creating a rupture that must be patched up in such a way as to preserve \textit{relevant global similarity} to the greatest possible degree. But which is the best way to patch things up? Are we really to believe that a world in which toppling kangaroos miraculously survive the crunch of natural selection is \textit{more similar} to ours than a world in which their bodies compensate in some way to regain their balance?\footnote{Snapping turtles, with their reduced shell on their underside (unlike painted turtles, which have full shells), have compensated by evolving a sharp beak, claws, and an aggressive disposition. This would seem to cast doubt on the counterfactual claim that “if painted turtles had a reduced lower shell, they would be rendered vulnerable.”} Yet prior to raising such
considerations, the intuitively obvious response is indeed that the counterfactual is true—they *would* topple over. In contrast, saying “if humans had no baby fingers, they would spurt blood from their hands” seems much less plausible, though it is not easy to say what makes the difference, much less to formalize the difference in an explicit standard. The lesson here is that the similarity relations between worlds, on which all assessments of all modally complex claims depend, are intimately sensitive to the particular claim being made, in ways that are subtle and difficult to discern.

Dretskean information, since it is stipulated to be distinct from mere *de facto* correlations, is inherently a modally complex phenomenon. For this reason, variations of Dretske’s theory, even if they avoid the problems that arise in dealing with channel conditions, will inevitably face difficulties of importantly the same sort. The “objective counterfactual theory” of Meskin and Cohen (2006), for instance, which takes the counterfactual *were the source not so-and-so, the signal would not be such-and-such* as definitive of the informational relationship, will succumb to all of the difficulties involved in evaluating counterfactuals. To see that these will be the same problems that confront Dretske, consider that Dretske’s channel conditions, since they support (at least certain cases of) this same counterfactual, can be understood as a roundabout way of providing a theory of its truth conditions. The modal complexity of information belies Dretske’s portrayal of information as “something the most reflective materialists should be willing to give,” the “physical yeast and flour” out of which we can build the mind (1981, p. xi)—physics, after all, is non-modal, and information is modally complex.

**Intelligence and Information**

Modal complexity, though confounding, is a natural and pervasive aspect of the human psychological world. The extent to which we experience the world as modally animated can sometimes make it difficult for us to discern the modal complexity in what we naïvely take to be the actual world. For instance, take Dretske’s comment that if there are Grebes “in the vicinity”, we do not know that we are seeing a Gadwell (1981a, p. 369). This is putatively a description of the *actual* world, on the basis of

---

9 Although Meskin and Cohen address some formal issues concerning counterfactuals, they do not discuss the more substantive concerns about the standards by which counterfactuals are evaluated, beyond asserting that they see “no serious reason to think the notion of counterfactuals is incoherent” (2006, p. 347). Coherence, needless to say, is not tantamount to objectivity.
which we can supposedly judge which alternative possibilities are relevant. But in fact “in the vicinity” is already a modally loaded notion; if someone is “in the area”, this just means, on a psychological level, that we might run into them. It is our naïve tendency to take the modally animated world of experience for the actual world that lends more plausibly to the idea of “the kind of possibilities that actually exist in the objective situation” (Dretske, 1981a, p. 377) than it deserves. If we think of actuality in terms of notions such as “in the vicinity”, as opposed to, say, “3010 metres away, north-by-north-east”, we will more readily accept the idea that the possibility that this is a Grebe right here is to be found “in the objective situation”.

In fact, the human ability to effortlessly perceive the world as modally animated might reasonably be taken to be a central and defining feature of intelligence itself. Douglas Hofstadter, for instance, has argued that “the crux of creativity resides in the ability to manufacture variations on a theme,” including “counterfactual conditionals”, “subjunctives”, and “‘almost’-situations” (i.e., near possibilities) (1985, p. 249). Consider the frame problem in artificial intelligence, characterized by Jerry Fodor as “to all intents and purposes… the problem of how the cognitive mind works” (1987, p. 148). The frame problem is the problem of predicting the relevant side effects of an event, while ignoring the irrelevant ones (see the discussion in Haugeland, 1985, p. 203-11). The problem is essentially that of finding the relevant true subjunctive conditionals that concern the immediate future, i.e., conditionals of the form if X happened now, Y would be the case. Although such conditionals are considered logically to be about the actual world (albeit in the future), the manner in which the future animates the present is psychologically similar to the manner in which nearby possible worlds animate the present in the cases of counterfactuals and, of course, relevant alternative possibilities. Zeroing in on relevant alternative possibilities, and bringing them to bear appropriately on the present situation, is something we humans are very good at—it plays a crucial role in our own assessments of the deliverances of perception. The problem of determining which possibilities are relevant, which, as we have seen, requires the appropriate application of past regularities to the idiosyncratic demands of the present situation, as well as the identification of the regularity and variability in the particular situation, the nuts and bolts and hinge points we perceive to be present in our modally animated world—this problem is at least as deserving as the frame problem of being considered central to intelligence. If there is any behavioral competence that is a sure sign of intelligence, it seems, to me at least, that the ability to discriminate relevant alternative possibilities is just as reasonable a
candidate as anything else we might think up. Like the frame problem, the problem of identifying relevant possibilities runs, in Fodor’s memorable phrase, “as deep as the analysis of rationality” (1987, p. 140).

Dretske endorses a view that he states in the form of a slogan: “If you can’t make one, you don’t understand how it works” (Dretske, 1994). This is an expression of Dretske’s view that the best test of a naturalistic theory of a mechanism is whether or not the theory tells us how to construct the mechanism. This view is put forward in much the same spirit as the Explicitness Condition, but it takes things one step further, telling us what counts as explicitness—if the theory of the mechanism is sufficiently explicit, enabling us to understand how it works, it should, ipso facto, tell us how we can build one. One of the virtues of Dretske’s buildability condition is that, like the Explicitness Condition, it serves the function of removing implicit, and (hence) potentially subjectivity-concealing elements: If our theory of how a mechanism works depends tacitly on unthematized subjective factors, these will surely be forced to the surface if we attempt to build the mechanism on the basis of the theory.

Could we, then, use Dretske’s buildability constraint to evaluate whether or not a standard serves as a suitable foundation for relational objectivity? This seems reasonable; if the standard is really explicit—that is, if it really provides a basis for making objective claims—we can, on the basis of the standard’s explicit formulation, construct a mechanism that serves the function of applying the standard. If we really understand how the rules of a game work, we ought to be able to build a mechanism that indicates whether or not the rules are being followed. What happens, though, when we try to build a mechanism that applies the standard of relevance that is the basis for the supposed objectivity of relevant possibilities, and, ultimately, of information? Of course, this is highly implausible, since, as we’ve observed, explicit standards of relevance are not so much as a gleam in the eye of any theorist today. But suppose we were able to construct a mechanism that could tell whether or not a possibility was relevant. If the preceding considerations are correct—those regarding the centrality of the recognition of relevant alternative possibilities to the nature of intelligence—then the mechanism we would end up building would be intelligent. The standard of relevance, naturally resistant to explicit formulation, will end up embodied implicitly in the thinking of a subject, one with interests and values, who pays attention to relevant possibilities, while ignoring irrelevant ones. The result for Dretske’s project is devastating, because he has got things exactly backwards: It is information that depends on minds, not the other way around.
Works Cited


Consciousness (Response to the Hard Problem)

Artem Tsvetkov

Cognitive Science and Biological Foundations of Behavior
Carleton University

Consciousness has an extensive history in human understanding of the self and the world. The word “consciousness” comes from Latin conscientia, literally meaning shared knowledge. Consciousness has been a research topic in fields of philosophy of mind, psychology, neuroscience, and cognitive science. In general, it can be described as a characteristic of the mind, including qualities such as subjectivity, self-awareness, and perception between the self and the world. It has been argued that John Locke was the first philosopher to define consciousness in modern terms in his famous Essay Concerning Human Understanding, although Rene Descartes was arguably the first philosopher to use it outside of the traditional scopes of understanding.

Among recent empirical findings on consciousness is an experimental investigation in which perception changes independently of the stimulus. Logothesis and his colleagues (Logothesis 1998) trained monkeys to pull different levers for different patterns, while monitoring the ongoing neural activity in monkeys’ visual system. When different patterns were presented to different eyes simultaneously, monkeys kept switching back and forth between the two levers even though the sensory input remained the same. In the lower visual areas 80% of the neurons did not shift with the percept, but further along the occipital-temporal pathway 90% shifted with the percept. This indicates that the occipital-temporal pathway (ventral stream) largely serves as the neural basis of visual consciousness. These findings were further extended and redefined using modern brain imaging technologies. Impressive correlations between neural activation and indications of perceptual experiences have been established by Kanwisher (2001). These neural correlates are all found in the ventral stream, signifying the identification of the neural basis of visual consciousness in the ventral stream.

An ongoing question for debate in empirical findings is if the difference between conscious and unconscious activation of the ventral pathway is just a matter of the degree of activation. Kanwisher (2001) proposes that, based on the evidence from ERP (Event-Related Potential) studies using the attentional blink paradigm, that neural activation of meaning is no less when the word is blinked than when it isn’t, suggesting that it is not lower neural activation strength that
accounts for lack of awareness. The ERP studies support the idea that the
difference between conscious and unconscious activation is a matter of neural
synchrony at fine timescales. Another idea is that the difference between seen and
“unseen” (seen but not attended to) stimuli might be a matter of interaction
between the classic visual stream and the areas of parietal and frontal cortex that
control attention.

Consciousness can be divided into access and phenomenal. Access
consciousness has to do with access of information in our minds for verbal
purposes, reasoning, or behavioral control. When we perceive, introspect, or
remember something it is all workings of Access consciousness. Phenomenal part
deals with experience itself – sensations, colors and colored forms, emotions and
feelings. In 1996 Chalmers formulated the Hard Problem, which deals with the
issue of “how to explain consciousness in terms of its neurological basis” (Block).
An easy problem, as opposed to the Hard Problem, would concern the function of
consciousness. The Hard Problem was identified by Nagel (1874) and further
analyzed in Levine (1983) (Block). Thomas Nagel proposed the idea that
consciousness and subjective experience cannot be reduced to brain activity. This
position is discussed in his most famous article “What is it like to be a bat?”
published in 1974 in The Philosophical Review. “I believe that there is a necessary
connection in both directions between the physical and the mental, but that it
cannot be discovered a priori” (Nagel).

There are various perspectives on consciousness, and particularly on the
Hard Problem. Among those are Eliminativism, Philosophical Reductionism
(Deflationism), Phenomenal Realism (Inflationism), and Dualistic Naturalism.
Eliminativism deals with the view that consciousness does not exist (Dennett 1979;
Rey 1997), thus there is nothing for the Hard Problem to be about. Deflationism
moves closer to common sense by allowing consciousness to exist and it can be
conceptually analyzed in non-phenomenal terms, based on behavior, functional,
representational, and cognitive analyses. Inflationism views consciousness as a
substantial property that cannot be conceptually reduced in non-phenomenal terms,
which is compatible with empirical scientific reduction to neurological or
computational properties of the brain. The Dualistic Naturalism view proposes that
there are naturalistic alternatives to Cartesian dualism – a deeper level of reality
that is the naturalistic basis of both consciousness and neuroscience.

The questions that the Hard Problem asks are: how could one property be
both phenomenal property and cortico-thalamic oscillation; and how is it possible
for something subjective to be something objective – first person vs. third person.
The Thalamus is a relay station between sensory input and the cerebral cortex. The
inputs are processed by the thalamus, and “converted” into a cortex readable
format. The strong bond that the thalamus shares with the cerebral cortex is
expressed by the thalamo-cortico-thalamic circuits, and it is these circuits that are believe to be involved in consciousness.

There are two points of view as to how consciousness relates to the brain. The first one is that a phenomenal property is a neural property; the second is that a phenomenal property has a neural basis. The first claim would be a Deflationist position, while the second is an Inflationist or a Dualistic Naturalist position. Essentially, the first claim says that there is nothing about consciousness that cannot be explained using non-phenomenal terms, for example – biological or neurological terms. The second claim is saying that there is a certain neural basis to consciousness, but that neural basis does not explain consciousness in its entirety. As stated by Block (2004), there are two reasons for thinking that the Hard Problem has no solution – actual and principled failure. The actual failure is that no one has been able to think of even a highly speculative answer to the Hard Problem; the principled failure is that the materials we have available seem ill suited to providing an answer. Nagel stated that an answer to the Hard Problem would seem to require an objective account that necessarily leaves out the subjectivity of what it is trying to explain. The claim that a phenomenal property is a neural property seems more mysterious than the claim that a phenomenal property has a certain neural basis. In his paper, The Problem of Consciousness, McGinn (1991) notes that neural phenomenal are spatial, although at first phenomenal appears to be non-spatial.

The rest of this paper is going to propose a new model for consciousness, attempting to explain how it interacts with the brain as well as provide a possible answer to the Hard Problem.

Arguably, every individual with a mind and a body exists in at least two realities – physical and mental. Physical reality is objective, while mental reality is subjective. The basis for stating that both, physical and mental, are realities is the fact that both influence one’s behavior as well as one’s thoughts. For example, damage to the brain or the spinal cord will result in severe alterations of one’s behavior, which signifies the physical reality. If one suffers from a mental degenerative disorder, such as Alzheimer’s disease, it will heavily affect the otherwise normal behavior, which signifies the reality of the mental world inside the physical world. Consider an individual that has undergone some very bad emotional experience. This experience will shape the mental reality, and further alter the behavior in order to prevent such an experience to occur again. This illustrates the reality of the physical world inside the mental world. While thinking about the experience, one may come to a deeper understanding of the matters that surround the situation, getting a greater insight that otherwise was not achievable via physical manipulations. Such mental manipulations signify the mental reality.
The relationship between the physical and the mental is reciprocal. From physical experiences we learn and obtain new sensations. These experiences are then analyzed by the mind, further facilitating understanding of matters involved in such an experience. Deeper and more profound understanding of a particular matter, especially a social one, results in alterations of behavior that can avoid a lot of potentially awkward or even dangerous situations. If a dangerous situation is avoided in the physical world, the result is once again mentally processed. Such an experience can be a rewarding one, as one was able to avoid danger by simply thinking of matters beforehand.

The mind can be divided into three components – conscious, subconscious, and unconscious. The conscious mind is the mind that executes various cognitive and self-regulating processes; subconscious is the support mind – purely mechanical and without awareness; the unconscious mind is simply the conscious mind that has been shut off. Conscious and subconscious are the two minds of interest in establishing this model, suggesting that the conscious mind resides in the pre-frontal cortex, while the subconscious mind is the rest of the brain.

The pre-frontal cortex is responsible for a wide variety of cognitive processes, such as thinking, decision making, and executive functions. The executive functions include differentiation among conflicting thoughts, good and bad, better and best, same and different, as well as consequences of actions and activities, prediction of outcomes and expectations, and socially-acceptable control. Depending on the degree of damage to the pre-frontal cortex, severe alterations to personality will occur. A classic case of Phineas Gage illustrates such an issue. Another example is a psychosurgical procedure called lobotomy. It consists of cutting the connections to and from, or simply destroying, the pre-frontal cortex. This procedure often results in major personality change, or even mental retardation, and was used in the past to treat such disorders as schizophrenia, clinical depression, and various anxiety disorders.

The rest of the brain, the subconscious mind, carries out mechanical procedures, such as remembering, and perceptual and motor functions. The parietal lobe allows us to maneuver our balance; occipital lobe allows us to see; temporal lobes store memories. Damage to the inferior parietal and frontal lobes has long been known to cause visual extinction in which subjects appear to lose subjective experience of certain stimuli on one side of the visual field. Kanwisher (2001) found that the very same ventral stream pathways are activated regardless of whether conscious processing is present or not. Stimulus of which the subject may not be aware is still being processed by the brain. This illustrates the “meaningless” (without consciously-attributed meaning) functional purposes of the subconscious mind.
Phenomenality is a functional property of consciousness. To draw an analogy, seeing is a functional property of the occipital lobe. Phenomenality, then, is the result of conscious processes that have a particular neural basis, just like seeing has a particular neural basis in the occipital lobe. It can be defined as a particular kind of organization or comparative process that includes inquisitive conceptual thinking, based on past experiences, to understand or “make sense” of the given input. Inquisitive conceptual thinking means thinking that includes application of known concepts to inquire deeper into the presented stimuli, in other words – to attribute a meaning to the situation. For example, if your friend is waving at you, the presented stimulus is the waving. It takes inquisitive conceptual thinking to realize that the concept of waving is a form of communication, from which you can inquire that your friend wants you to come over. Phenomenality intertwines with epistemology – the study of knowledge. Just as knowledge can be defined in neurological terms, suggesting that there is some storage of memories and concepts that are logically traversed to make inquiries, phenomenality can be defined neurologically by examining the neural structures in the pre-frontal cortex. This neurological definition, however, will not give an insight into the abstract understanding and meaning carried out by phenomenality, thus including only the physical reality component, but leaving out the mental reality one.

To create an abstract analogy that will illustrate the mental component, spatial properties of physical reality can be considered. The physical reality consists of three dimensions – x, y, and z. Each of these dimensions is at a 90 degree angle to each other one. To follow through with the analogy, the mental reality can be said to be another spatial dimension, which is also at 90 degrees to x, y, and z. This dimension is no longer a physical one, but a mental – parallel to physical – one. It is a subjective dimension, while the other three are objective. Just as a physical experience contains a spatial or a visual component, a mental experience may contain those as well. As a response to McGinn’s statement that phenomenal is non-spatial, I state that phenomenal, in fact, can be spatial, as well as visual and verbal. REM (Rapid-Eye Movement) during sleeping has been found to correlate with the experience of a dream. Dreams often contain spatial experiences of being somewhere, visual experiences of seeing people and objects, and verbal or otherwise communicative experiences. There is a neural basis for the phenomenality of these experiences, but it does not account for the meaning. In this, the answer to the first question of the Hard Problem emerges.

The neural activity, expressed as a cortico-thalamic oscillation, is just that in the physical reality, while the phenomenality of it is the subjective meaning that occurs in the mental reality. Both processes are parallel and are in conjunction with each other. The Thalamus, then, can be said to be a physical link between the two realities, implying that there is a mental link present as well. The study of the
mental link and the mental reality includes spiritual practices, such as Buddhism and shamanism. Spirit itself is another word for mind. Spirituality, then, is a systematic study of the mental reality, just as science is a systematic study of the physical reality.

Computationally, this model can be seen as computer hardware – the physical – and computer software – the mental. On a physical – neurological – level we see hardware; magnets and electrodes; chipsets and silicon. However, by simply looking at a particular computer’s hardware we can draw no conclusions about what is in the mind of that computer. We obtain no insight into the phenomenality that lies beyond. The phenomenality itself is expressed by the operating system and all the programs that run on it; virtual reality; internet. As in the analogy above, it can be said that the virtual reality spatially exists at a 90 degree angle to the physical reality.

To answer the second question of the Hard Problem, first- and third- person views need to be defined in terms of the subjective and objective. A first-person experience consists of a subjective interpretation or experience of some given events. The experience is personal, and means different things to different people, completely depending on individual phenomenality. In a third-person view there is an intermediate step between the first-person subjective interpretation of some events and the events themselves. The intermediate step is an outside body, or simply put – another individual, that possesses his or her own phenomenality. Third-person view is experienced when events are interpreted by an outside mind, followed by the interpretation being communicated verbally or otherwise. Thus the result, in the first-person view, is an interpretation of an interpretation. In terms of the subjective an objective, the definitions are as follows. First-person view is a subjective experience of an event; third-person view is an objective experience of a subjective experience of an event. To further illustrate the definition, consider an example:

I am looking at a table. My phenomenality subjectively interprets the object. It is my table, so all the past memories of me writing papers, playing computer games, and recording music are associated with it. I am receiving a first-person experience of the table. As you are reading this, you are experiencing my table in third-person. Your phenomenality is receiving an objective experience of my subjective experience with the table, followed by your personal subjective interpretation of what that table may be like. The above example illustrates how something first-personal can be something third-personal, thus answering the second question of the Hard Problem.

An inflationist dualistic naturalist position that this model takes is able to coherently illustrate the solution to both question of the Hard Problem, putting itself into the “highly-speculative answer” category. The mind is not a mysterious
matter; it is simply outside of the physical understanding. Buddhist monks have been burned alive while meditating by Mongolian invasions of Tibet; these same monks have been known to achieve levitation. Shamans are known as communicators with the spirit world, and as already established – spirit is the mind. Spiritual practices such as Buddhism and shamanism are merely mind exploration practices, which transcend the limits of the untrained mind. As stated by Nagel – there is a connection between the physical and the mental, but it cannot be known a priori. And indeed – the connection must be experienced and understood subjectively before it can be known.

References

The Joining of Neuroscience, Psychology, and Philosophy in a Search for the Self

Samantha Noll

Philosophy and Anthropology
West Chester University of Pennsylvania

“Who are you?” seems like a question with a simple answer. When asked, a person often replies automatically with their name. When further pressed, they may supply their place of origin or what they do for a living. When pressed even further, often they shrug their shoulders and ask in reply, “What do you want to know?” The seemingly simple question, “Who are you?” is no longer able to be answered with a rote answer; with further pressing the question reveals its complexity. Is the self bound up with our personal history, our narrative, or is it more than that? What makes us who we are? Is the self to be found in the intricate system of neurons in the brain or is it something that is separate from our biological bodies? Are our thoughts, desires, pains and all other mental states dependent solely on the roles they play in the cognitive system (functionalism) or is it better to focus on the four features of the mind (consciousness, intentionality, subjectivity, and mental causation) in order to understand what the self is? This paper puts forth that none of these methods will give us a full answer to our question. However, by incorporating insights found within many different theories of the mind, we will be able to move closer to an understanding of the self. In order to understand what makes us who we are, we need to look at the pattern of the system, or systems, and not just the individual parts.

First, we need to define our terms before moving on. The definition of the word “self” found in The Harper Collins Dictionary of Philosophy includes an interesting entry that reads, “[The self is] the unity (ego, subject, memory, mind, I, awareness, consciousness-knower) that endures throughout change and is aware of its unity, its endurance, and the change” (Angeles 269). Another interesting entry found in the dictionary under “self” reads as follows, “[The self is] the entire sequence of mental events of which one can be aware at a given moment”

(Angeles 269).\(^2\) Now, the definition for “mind” is more drawn out and begins as follows,

**Mind**

1. consciousness; awareness.
2. human rational powers; thought; the capacity to think.
3. psyche; self; ego; personal identity.
4. soul; spirit; spiritual substance.
5. that which endures throughout changes of consciousness (experience, awareness).
6. the entity that performs such functions as sensing, perceiving, remembering, imagining, conceiving, feeling, emoting, willing, reasoning, extrapolating into the future, or judging…\(^3\)

Only about half of the definition found in the dictionary for “mind” is quoted because it continues at length and the aspects important to answering our question are to be found above. Both the definition for “mind” and for “self” are very broad; they both include features and functions and they both include aspects of each other. For the purposes of understanding the self, it is important to note that self and personal identity are included in the definition of “mind” and both mental events and mind are included in the definition of “self”. Therefore, we can safely assume that the system that makes up the self is intricately bound up with the mind and vice versa. I would even venture to say that they are different aspects of the same system. This is not a view taken on purely semantic authority; it is also the view that the neuroscientist Susan Greenfield takes in her book, *The Private Life of the Brain*. She writes, “After all, if mind is the personalization of the brain, then what more, or what less, could self actually be?” (Greenfield 186).\(^4\)

If the self if bound up with the mind then this brings us to the mind-body question. Today, several theories of mind, from several different fields, focus on the biological aspect of what makes us who we are. Just peruse the shelves of your local book store and you will find a dozen titles that suggest just that. For example, two titles that I’ve spotted with a quick glance at my own library include: *Astonishing Hypothesis: the Scientific Search for the Soul* by Francis Crick and *A Universe of Consciousness: How Matter Becomes Imagination* by Gerald Edelman. What is so interesting is the rather recent acceptance of a purely materialistic approach to understanding the self. The ascendancy of monism and, in particular, physicalism in the 20th and 21st centuries has led to such positions in


philosophy of mind as behaviorism, the identity theory, and functionalism.\(^5\) In fact, in the *Stanford Encyclopedia*’s entry for physicalism, Stoljar writes, “Most contemporary philosophers are physicalists” (Stoljar).\(^6\)

However, prior to the 19th century, many philosophers who wrote about the mind-body question where considered dualists (such as Plato and Descartes). According to Howard Robinson, “In the philosophy of mind, dualism is the theory that the mental and the physical — or mind and body or mind and brain — are, in some sense, radically different kinds of things” (Robinson).\(^7\) With the publication of *The Origin of Species* by Charles Darwin and the rise of modern science, the tide began to shift for dualism. At that time, a more mechanistic view came into favor. Mechanism is, roughly, the idea that all natural phenomena “follows from and is in accord with the laws of physics” (Robinson).\(^8\)

In the early 20th century, largely stemming from the need to answer the problems brought about by mechanism, different forms of material monism were devised. But, according to Howard Robinson, even though “dualism has been out of fashion in psychology since the advent of behaviorism (Watson (1913)) and in philosophy since Ryle (1949), the argument is by no means over” (Robinson).\(^9\) It is my view that the breakthroughs in science can and will move us further towards our goal of understanding what makes us who we are. However, it is only part of the picture. In order to fully understand the self, we must first look at a few philosophers from both the dualist and the monist traditions that have made major contributions to the philosophy of mind.

In the *Treatise of Man*, Descartes did not describe humans but, rather, he described conceptual models of humans that consisted of two ingredients: a body and a soul. He regarded the soul as “the principle of thought” and believed that it was to be found in a specific area of the brain known as the pineal gland.


For Descartes, the body has a life of its own but is still connected to the soul. This soul acts on the pineal gland and produces such mental events as sensing, perceiving, willing, and thinking (Angeles 76). He wrote, “Since it is the only solid part in the whole brain which is single, it must necessarily be the seat of the common sense, i.e., of thought, and consequently of the soul” (qtd. in Lokhorst). Prior to this, St. Augustine wrote in On the Trinity, “In each body the whole soul is in the whole body, and whole in each part of it” (qtd. in Lokhorst). Descartes moved away from the idea of the soul being found in the whole body and positioned the seat of the soul in the brain.

It is interesting that Descartes was both a substance dualist and a mechanist. Also, his definition of soul is more closely aligned to the modern definition of mind. With the publication of the Treatise of Man in 1632 (and later in the Passions of the Soul published in 1649), Descartes wrote of a soul-body dualism, which has today turned into mind-body dualism and, even though he was wrong about the importance of the pineal gland, he pinpointed the part of the body (the brain) which is responsible for thought. Later Philosophers, such as Hume, would pick up on this idea and expand upon it.

Hume criticized Descartes’ idea of substance dualism because he found the whole concept lacking in empirical content. I.e. “when you search for the owner of the properties that make up a substance, you find nothing but further properties” (Robinson). Hume thought that the mind is “nothing but a bundle or collection of different perceptions, which succeed each other with an inconceivable rapidity, and are in perpetual flux and movement” (qtd. in Angeles 188).

---


separate from the brain but is a collection of experiences that occur between birth and death. The entire series of these events is called the bundle and this bundle is the mind or self. The events that make up the bundle “are related by such features as: (a) resemblances of perception, (b) contiguity of experiences in time and place, (c) regularity of succession among perceptions, and (d) memory” (Angeles 188). If any of these components are missing, then it cannot be said that you have a mind and, for Hume, the mind does not exist apart from the brain.

Hume’s view is known as bundle dualism but it is not necessarily dualism because it is a theory about the makeup of the unity of the mind; it has also been accepted by some physicalists including Parfit and Shoemaker (Robinson). While I do not agree that a collection of experiences is all that makes up the mind, I do think that Hume contributes two important pieces of the puzzle that will bring us closer to our goal of understanding the development of self. First, Hume separates different features of the mind (such as memory and perception) which play important roles in the creation of the self. And second, he also stressed the role of space and time in the ordering of experiences, which is of paramount importance in the creation of the self. Also, we will see later that the prefrontal cortex of the brain and the hippocampus (found in the medial temporal lobe) have been linked to the creation of memories that include space and time.

Over two hundred years after the death of David Hume, it is now thought by many philosophers and scientists that the unity of the mind, which Hume was searching for, is to be found within the systems of the brain itself. Both Descartes and Hume agreed that the mind was to be found within, or could not be separated from, the brain (for Descartes it acted upon the pineal gland specifically). However, Hume recognized that the experiences which occur between birth and death also make up the mind or self. These experiences are bound by features (such as memory) which are all aspects of the system, or systems, found within the brain.

John Searle, in an essay entitled The Mind-Body Problem, argues that all mental phenomena (both conscious and unconscious) are caused by processes going on within the brain. He writes, “to put it crudely, and counting all of the central nervous system as part of the brain for our present discussion, everything that matters for our mental life, all of our thoughts and feelings, are caused by processes inside the brain” (Searle 19).

failure to solve the mind-body problem is a misunderstanding of causation.\textsuperscript{20} Hume used the model of two billiard balls hitting each other to represent cause and effect, but for Searle, this is a crude understanding of causation, especially when applied to the relationship between the mind and the brain. Searle turns to the common distinction between micro- and macro- properties of systems found in physics in order to gain a greater understanding of the mind-body relationship.

Every object, like a desk or pen for example, is made up of micro-particles. These micro-particles have structure and features at the level of atoms and molecules and, if you go deeper, they also have features at the subatomic level. In addition, objects have surface or global features; an example of a surface feature is the solidity of the desk mentioned above. The desk has features at both the atomic and surface level. The surface features are due to the behavior of elements at the micro level. Searle states, “… macro features are causally explained by the behavior of elements at the micro-level” (Searle 21).\textsuperscript{21} He uses examples throughout his paper (such as a table, water, and a hammer) in order to further explain his point. All of these objects, and indeed, all of the objects around you, have both micro- and macro- properties. Searle states that, “Nothing is more common in nature then for surface features of a phenomenon to be both caused by and realized in a micro-structure, and those are exactly the relationships that are exhibited by the relation of mind to brain” (Searle 22-23).\textsuperscript{22} Therefore, in order to understand the relationship between the mind and body, we must have a better grasp of the processes that make up the system.

For Searle, mental states are features of the brain that can be described at two levels, a macro-level (or higher level) in mental terms and a micro-level (or lower level) in physiological terms. “…The mind and body interact but they are not two different things since mental phenomena are features of the brain” (Searle 26).\textsuperscript{23} This goes beyond simple cause and effect because both the mind and the physiological features (such as neuronal processes) are part of the same system. With this view, both the physical phenomena and mental phenomena exist; it is a bridging of both native physicalism and native mentalism and it brings us one step closer to understanding the self.

Searle’s theory shows how closely the mental and physical levels are bound together. The mind, and indeed the self, is no longer the inessential froth on a wave


(as mechanists conjectured) nor is it something that is immaterial. It is now intricately bound up with the body and, as we will discuss later, with the physical world around us. If the self is a part of the system, then we will be able to understand the processes that are essential in its creation. For, if the mind and brain are different levels of the same system, then the events that we experience between birth and death actually contribute to the development of the brain. This is a view that many psychologists and neuroscientists have taken up in recent years.

One neuroscientist that takes up this basic premise in his book *Synaptic Self: How Our Brains Become Who We Are* is Joseph LeDoux. He draws upon massive amounts neurological and psychological research to support his theory that “the particular pattern of an individual’s brain, and the information encoded by these connections are the keys to who that person is” (LeDoux 3). For LeDoux, the self, your personality, who you are, is to be found in the patterns of connections between neurons, which are known as synapses. These synapses “are the main channels of information flow and storage in the brain” (LeDoux 2).

LeDoux realizes that many people view the self as “psychological, social, moral, aesthetic, or spiritual” but he states that his theory is not an alternative to these views, it is just an attempt to portray how that “psychological, social, moral, aesthetic, or spiritual” self is realized (LeDoux 2-3). “People don’t come preassembled but are glued together by life” (LeDoux 3). The first step to understanding what makes us who we are, is to realize that both nature and nurture contribute to the shaping of the self. In fact, nature and nurture speak the same language and synapses are the key to understanding both (LeDoux 3-5).

For years there has been a “nature vs. nurture” debate but now it is generally accepted that both play important roles. It is presently “not so much a debate about genes vs. environmental experience as one about the precise contribution of

---


experience” (LeDoux 72). Roughly speaking, “genes only shape the broad outline of mental and behavioral functions, accounting for at most 50 percent of a given trait, and in many instances far less” (LeDoux 5). Our genes may bias our development in particular directions but many other factors contribute to how our genes are expressed.

For instance, a child with a tendency towards shyness could, instead, overcome that shyness because he or she had supportive and encouraging parents. Likewise, a child with a tendency towards sociability could become introverted because of an unsupportive family environment while growing up. LeDoux points out that, “Most systems of the brain are plastic, that is, modifiable by experience, which means that the synapses involved are changed by experience” (LeDoux 8). The self may be biased towards developing in a particular way, but it is ultimately the effects of the environment that shapes how traits will be displayed.

What is interesting is that the plasticity of the brain is determined by our genes. The framework that the self is woven upon (the basic makeup of our brain) is determined by our genes but not the individual self that each of us becomes. Each individual is unique and one of a kind because each person is a product of their genes as well as their environment and experiences. For LeDoux, without learning, and its synaptic result, memory, “a person would have a bare-bone personality provided by genes, but wouldn’t know much about it… [Learning and memory] play major roles in gluing a coherent personality together as one goes through life” (LeDoux 9).

Now, in order to fully understand the role that learning and memory play in the development of the self, we must further define our terms and look at the different types of memory outlined by LeDoux in *Synaptic Self*. Prior to the 1980’s the term “memory” was used to describe both explicit (facts and experiences) and implicit (conditioning, skills, and priming) forms of long-term memory. But because of the work done by Larry Squire and Neal Cohen (who originally used the terms declarative and procedural memory in 1980) and because of the work done by Dan Schacter (who specifically coined the terms explicit and implicit

---


memory in 1986) the distinction between the different forms of memory was made.  

Explicit memories are memories of things that we were once aware of; they are memories that we can consciously recall, that we have access to. We can remember events, people, scenes, songs, paintings, and any number of things that we were once aware of. This is what most people think of when they think of the term “memory”. Implicit memory is memory that we have limited or no conscious access to. Implicit forms of memory contribute in very important ways to the development of the self; it contributes towards our personality traits. “Each of us has his or her own style of walking, talking, and thinking. We hold our bodies in a certain manner when we are standing or sitting… these and many other aspects of behavior are expressed so automatically, so implicitly, that they may seem unchangeable, perhaps innate” (LeDoux 117). 34 I agree with LeDoux that we should not overlook the role of experience, of learning and memory, in establishing these aspects of our behavior and in maintaining them.

So, in essence, we have two layers of memory that make up who we are: we have the explicit memory of our experiences and facts (which we can consciously mold and revise as time goes on) and the implicit memory that makes up our conditioning, skills and priming. “Through explicit systems, we try to willfully dictate who we are, and how we behave. But we are only partially effective in doing so since we have imperfect access to emotional [implicit] systems, which play such a crucial role in coordinating learning by other systems” (LeDoux 323). 35

Explicit memories, as Bartlett demonstrated in 1932, are not stored as complete entities within our brains; rather they are constructions assembled at the time of retrieval. The information stored during the original experience is only one of the items used to construct the memory. Other contributing factors to your memories include information already stored, how you feel about the event, and what you hear or see after the experience (LeDoux 203). 36 In fact, studies by Elizabeth Loftus and others, in 1986 and 1989, show that memories of particularly

---


emotional events are often much different than the actual events (LeDoux 203). It is not that the recollections of these events have been intentionally fabricated, but rather remembered experiences (explicit memories) are impressions of the actual experience.

This idea falls right in line with Ricoeur’s concept of narrative identity. For Ricoeur, the self can be viewed as the principle character in a narrative. “If in living my life I configure it as a narrative, I understand my life by refiguring it: ‘the fragile offshoot issuing from the union of history and fiction is the assignment to an individual or a community of a specific identity that we can call their narrative identity’” (Simms 102). However, in order to fully understand identity, Ricoeur proposed that we must divide it into two parts, ipse identity and idem identity.

Ipse identity is narrative identity, the identity that you have control over; the identity that you can sculpt and form into a coherent narrative of your life. This is the part of identity that you have conscious access to. Narrative identity is developed when memories are consciously shaped and formed into a narrative structure and it is through this process of “writing” your narrative, that you come to understand your life. Thus, it is formed by your explicit memories. Idem identity is the identity of the body; it is the identity of “sameness” through change. Everyone ages and changes over time but, even though a person is older, they are the same person bodily; they are the same but changed by age and this “sameness” is idem identity (Simms 102).

The memories that you store in your brain are constructed at the time of retrieval; memories are constantly changing and being rewritten. This process of rewriting and organizing your past experiences into a coherent whole is how you create your narrative identity and you are constantly rewriting that narrative. This is a major part of the creation of the self but it is only one part of what makes you who you are. Implicit memory and emotion also play an important role in the self. This is because implicit memory and emotion shape and skew how you view certain events and people, as we will fully explore later. But first, in order to understand the role that implicit memory and emotion plays in the creation of the self, we must look at a few developments in both philosophy and neuroscience.


Two views of the self prominent in contemporary philosophy include the Kantian view of the self as an ethical subject that “uses reason to transcend cultural norms and to discover absolute moral truth” and the utilitarian view of the self or homo-economicus; a self that uses reason to rank desires and to develop a strategy that will maximize the satisfaction of those desires (Meyers). However, both of these views downplay the importance of emotion and this privileging of reason over emotion has been a dominant theme in western philosophy. *Cogito ergo sum*, being just one example out of many.

It is interesting to note that until recently, the importance of emotion was mostly ignored and relatively unexplored by neuroscientists as well. In fact, by the mid 1960’s (after the development of the limbic system theory of emotion by Paul Maclean) neuroscientists predominantly viewed emotion as “more a manner of mental content than of mental processing” and thus not relevant to their field (LeDoux 201). Because of the subjective nature of emotion, it was difficult to measure and study and, until recently, the important role of emotion in the self was not only downplayed in philosophy, as it has been for centuries, but also in neuroscience, a discipline dedicated to understanding the human brain and mind (LeDoux 201-202).

However, a recent approach to studying emotion called *emotional processing* has once again made emotion a popular topic in neuroscience. Emotional processing does not concentrate on the subjective states of emotion; rather it is the study of the way these emotions come about. “Specifically, from this point of view, emotion can be defined as the process by which the brain determines or computes the value of a stimulus. Other aspects of emotion then follow from this computation” (LeDoux 206). This approach further defines the term “emotion”. Historically, it has been understood as meaning “feeling” but a more exact definition should include the unconscious processes of emotion as well as the feelings that these processes produce. In this way, we can further understand the role of emotion without relying strictly on subjective experience or behaviorist

---


models. In the *Webster’s Dictionary*, the first entry under the word “emotion” reads, “strong feeling” (*Webster’s Dictionary* 195). However, the reactions to the processes going on within your body, that can usually be recognized and named by your conscious self, are only one part of what makes up emotion.

For example, the first thing that happens when a stimulus, that will produce an emotional response, is detected by the brain (such as a car flying by or a ball thrown at your head) is an emotional reaction, or reactions, such as an increased heart rate. Subsequently, “a feeling emerges as we become aware that our brain has determined that something important is present and we are reacting to it” (LeDoux 206). For LeDoux, it is a relatively easy thing to account for how emotional reactions can follow from emotional processing. “Information received by sensory systems activates emotional-processing circuits [a circuit is a group of neurons linked together that performs some specific function], which evaluate the meaning of the stimulus input and initiate specific emotional responses by triggering output and only afterward noticed what it was- a ball thrown… for example” (LeDoux 207-208).

In the case of a ball thrown at your head or a car flying by, you only notice the feeling of fear after you jump back and after your heart is already pumping. The feeling itself did not cause the jumping or the increase in heart rate. There are many more examples given and studies performed that have proven these facts (See pages 206-210 in *Synaptic Self*) but the important thing to note is that the study of emotional processing has shown the fundamental nature and role of emotion in everyday life and in the development of the self, as we will explore later. “A Processing approach… allows emotion and cognition to be treated the same… and opens the door for the much needed integration of cognition and emotion” (LeDoux 209-210).

Another important fact to note is that emotional conditioning is a form of implicit learning. Subjects can be conditioned to have implicit reactions to stimuli. This fact, known as classical conditioning, has been understood since the time of Pavlov and his dog. For example, in the case of Ivan Pavlov, roughly, if you ring a bell while feeding a dog, eventually the sound of the bell will be associated with

---


food and, from that point on, the dog will salivate at the ringing of a bell. In classical conditioning, a subject learns to associate two stimuli. It is now thought that circuits that do not depend on the explicit memory system instead engage in implicit learning (LeDoux 119).48

It is important to note, though, that today neuroscientists “have a detailed understanding of circuitry involved in only a few of the many systems that learn implicitly” (LeDoux 119).49 This is an area in neuroscience where there is still a great deal of research to be done. However, we should not underestimate the role that implicit memory has in the creation of the self. Studies done by Liz Philips and Paul Whalen have shown that implicit memory is activated in social interactions. “In separate studies, they found that exposure of white subjects to the faces of unfamiliar African Americans led to amygdala activation (a part of the brain thought to be associated with emotion), and the degree of activation was directly related to the subject’s score on a test that measures racial biases” (LeDoux 221).50 The bias test shows that people can be conditioned to have implicit tendencies towards racism.

Implicit memory affects you in more ways than just the way you walk or the tones of your voice. It is very important to realize the power that implicit memories have on the type of person you are and on society as a whole. It is also important to recognize and, therefore, to be able to begin to change, tendencies towards behavior that is undesirable and unacceptable. These changes cannot be accomplished if we refuse to see the power of implicit memory. It is an essential part of what makes us who we are and it too has been shaped by experience and our environment.

In the past, philosophy has predominantly focused on explicit (or executive) functions within the brain (such as cognition) and has ignored the implicit systems within the brain (such as emotion and the role of implicit memory). This focus on the strictly executive functions of the mind has created a skewed image of the self. Emotions play an important part in our lives; they help us to stay alive, to remain healthy, to identity safe environments, to propagate our species, to detect friend from foe, to bond with others and form communities, and many other functions. Therefore, we are not only rational beings but also semiotic beings. These two


layers of self are intertwined, equally important, and inseparable. Historically, emotion was set up as the antithesis of reason. However, reason, as modern findings in neuroscience have made clear, is only one of the aspects that make us who we are; it is not the antithesis of emotion, but rather both emotion and reason perform important functions. They work in unison for the survival of the whole. If we focus on just one aspect of what makes us human, of what makes us who we are, then we will invariably cultivate (and have cultivated) a skewed view of reality, humanity, and the self.

The intertwining of emotions and cognition leads us to the work of Julia Kristeva. For Kristeva, the self includes both the semiotic and the symbolic. “Kristeva understands the self as a dynamic interplay between the feminine semiotic and the masculine symbolic…” (Meyers). Unlike the symbolic, the semiotic gives expression to repressed and unconscious material. It is the body; it is emotion, intuition, and feeling. The semiotic is implicit memory. Kristeva understands the lessons that language can teach us about the self. She understands that the self is more than just the symbolic; it is more than alienated reason and explicit functions. We think but we also feel, dream, create, desire, and intuitively understand. The self is dynamic and Kristeva’s view falls in line with the findings of modern neurological research and her view of the semiotic is supported by our increasing understanding of the role that implicit memory plays in the creation of the self.

“For Kristeva, the self is a subject of enunciation — a speaker who can use the pronoun ‘I’” (Meyers). And both the symbolic (characterized by signs and linear logic) and the semiotic (which corresponds to the unconscious and emotions) can be found in language. This falls in line with the view of the linguist Ray Jackendoff in his book Foundations of Language. He states as a basic principle in his book, “that the proper formulation of reference [in language] is as a relation between linguistic expressions and the world as conceptualized by the language user” (Jackendoff xvi). If language is the linguistic expression of the world as seen by the language user, then it must include both the symbolic and the semiotic because both of these systems are involved in how we see and understand the

---


world around us. Thus, language is an intertwining of the explicit and implicit systems and, in that way, is an act of creation that reflects the creation of the self.

Now, it is interesting to note that language has greatly increased the power of the brain. This is because it gives us the ability to “chunk” or categorize large amounts of information (LeDoux 177). Working memory is involved in all thought and problem solving and is one of the brain’s most sophisticated capacities. It underlies the ability to read a menu, to play chess, or to have a conversation (LeDoux 175-176). However, working memory can only keep a few things active in the brain at one time and the ability to chunk increases the amount of information that you can deal with at once. For LeDoux, “It is... the structuring of cognition around language that confers on the human brain its unique qualities” (LeDoux 197). There is a big difference between having only a nonverbal working memory and having both a nonverbal and a verbal working memory. “Language radically alters the brain’s ability to compare, contrast, discriminate, and associate on-line, in real time, and to use such information to guide thinking and problem-solving” (LeDoux 197).

It is language that makes it possible for us to signify “I” and, thus, to be a subject of enunciation. This is because complex abstract concepts are represented by language (such as “I”, “me”, “ours”); we relate external events to these abstractions and use them to guide decision making and thought. Language is personal but it is also public because the signs that make up language constitute groupings of information that have been accepted by the culture that each of us is a part of. These groupings mold and direct thought because complex thought is dependent upon language to chunk information. In this way, Wittgenstein was correct when he argued against private language in *Philosophical Investigations* because our thoughts are necessarily shaped by the language that we use. A person can not have a purely personal language that is only understood by them (if you view personal language as “language”) (Candlish).

---


However, language is still used to signify the world “as conceptualized by the language user“, and is only one system within the brain; we have both verbal and nonverbal working memory capacities (Jackendoff xvi). Our minds, and our selves, are shaped by language but language is also shaped by the mind. Thus, Kristeva’s view that language contains both the symbolic and the semiotic shows that language is a reflection of the structure of the mind which contains both explicit and implicit systems and is shaped by both nature and nurture. Our explicit memories are dependent upon what we have learned before. “Indeed, much of the self is learned by making new memories out of old ones. Just as learning is the process of creating memories, the memories created are dependent on things we’ve learned before” (LeDoux 96).

If the structure of language is a reflection of the structure of the mind and therefore of the self, then we can utilize theories concerning language in order to understand the self. As we explored earlier, a part of what makes us who we are is our narrative identity. According to the work done in the 1970’s by the psychologist Endel Tulving, long-term memories can be broken down by what they are about. He argued that there are two categories of explicit memories: episodic memories dependent upon space and time (things that happened to you) and semantic memories which are simple facts that are stored rather than experience (LeDoux 108). Recently, the work of Faraneh Vargha-Khadem, Mortimer Mishkin, and others supports Tulving’s distinction. They studied children who damaged their hippocampus early in life. Even with their injuries, they were able to attend school and learn basic facts despite having a poor ability to remember their own experiences. Many researchers think that the hippocampus is specifically involved in remembering episodic memories (LeDoux 108).

Your narrative identity, based upon episodic memory, is how you come to understand yourself and your life. It has a simple plot structure, moving from earlier events to a climax, which is ultimately death. Language also moves from earlier words to later, building meaning with each word. I propose that it is not the self that is constructed like a narrative but, rather, the narrative that is constructed


like the self. If language is a reflection of the structure of the mind, then we could infer that the constructions of language (i.e. narratives) also share that structure. Remember, the self is not completely developed by outside experiences but, rather, it is the dynamic interplay between environment and genes; just as narrative is not completely developed by the symbolic but is the interplay of both the semiotic and the symbolic. Narrative identity, therefore, is the dynamic interplay of nature and nurture, of the outside world and the inside world, and both worlds are affected by the other. The narrative is shaped by the mind just as the mind is shaped by language.

If part of our identity is a narrative, then we can apply Derrida’s idea of differance in order to understand how the narrative self gains meaning. Every experience is temporal; the moment of the present, of now, is different from the past and future, but it is a small difference that is non-dualistic. This difference is undecidable (Lawlor “Jacques Derrida”).63 The cutting of reality into dualisms is an artificial partitioning of what is. Each concept that makes up part of a dualism depends upon the other concept for meaning. One cannot be valued over the other because each makes the other possible; each contains the other; each is within the other. Deconstruction and, therefore, the principle of differance is not only a textual strategy, a particular technique that can be used to unpack texts, but rather, it is the event; it is what is found in appearances when we realize that they are temporal. It is what lies beneath the cutting of reality into dualistic concepts. The French word differer means both to “defer” and to “differ”. The difference between words makes meaning possible and, at the same time, constantly defers meaning. For Derrida, “Literature, like thought, cannot be reduced down to either a theory of imitation (imitation of actual things or of the idea) or to a theory of creativity. Instead, literature ‘is’ the simultaneity of imitation and creation, of identity and difference…” (Lawlor, Imagination 117-118).64

Derrida built upon the work of Saussure in order to explain how texts gain meaning through differance. For Saussure, a sign is made up of two components: the signifier, which is the actual word, and the signified, which is the concept that the word represents (Saussure 115).65 Signs are arbitrary; but then Saussure goes on to explain that signs are also differential. Saussure writes, “It is precisely


because two signs $a$ and $b$ are never grasped as such by our linguistic consciousness, but only the difference between $a$ and $b$, that each sign remains free to change in accordance with laws quite unconnected with their signifying function… all that matters is the difference between the signs…” (Saussure 116).

Derrida agreed with Saussure that without the spaces between signs, there would be no stationary meaning; this is because the sign’s meaning can change in accordance with the rules of the linguistic system. The meaning of a sign is contingent upon the linguistic system which is made up of differences. We do not grasp the meaning of the sign itself, but rather we understand the sign because of its place in the system, because of the signs that follow. In the essay entitled *Differance*, Derrida wrote, “Every concept is necessarily and essentially inscribed in a chain or a system, within which it refers to another and to another and to other concepts, by the systematic play of differences” (Derrida 449).

In the same way that the meaning of a word is constantly deferred down an endless chain of signifiers, the meaning of an explicit memory is constantly deferred down an endless chain of memories; this is because our understanding of events is contingent upon new events, pieces of information, changing feelings, and past information both collected and forgotten. How many of us have looked back upon an event that we once thought had a particular meaning, only to realize that, as we gained in experience, the event held a completely different meaning. The mind utilizes a system in order to organize memories. It forms relations and configurations of stimuli, of memories about spatial arrangements, and among different memories; it is able to join together many pieces of information at once and to configure them into a context (LeDoux 132).

The mind does this automatically and without the partitioning off and organizing of memories there would be no meaning and, therefore, no self. Just like with words, the difference between memories makes meaning possible and, at the same time, constantly defers meaning. Identity cannot be privileged over difference and presence cannot be privileged over absence because each concept makes the other possible. Your identity and the world around you could not be understood without the spaces; without a system made up of differences and the partitioning of reality into manageable chunks.

---


A word signifies a breaking away, a partitioning off of reality. In order for a sign to have meaning, it must be seen against the backdrop of other signs not present. We know light because of dark, future because of past, joy because of sorrow, etc. But this breaking into pieces, this partitioning off, and this forming of dualisms, is both how we come to understand reality and an artificial structure built upon reality; it is not reality itself, but rather a system that enables us to understand and to communicate. This difference, thus, gives meaning to words and concepts but only against the backdrop of other words and concepts not present. Derrida wrote, “We could thus take up all the coupled oppositions on which philosophy is constructed, and from which our language lives, not in order to see opposition vanish but to see the emergence of a necessary such that one of the terms appears as the difference of the other, the other as ‘differed’ within the systematic ordering of the same…” (Derrida 455).

Language helps us to recognize the difference that makes all meaning possible. The structure of language points to the structure of the mind and the mind automatically recognizes and separates space and time; it breaks down the outside world into “chunks” that can be handled by the capacities of the brain. It creates binary oppositions but does not automatically privilege one over the other. We are bound by the limits of the brain and by the senses that we are born with. Language points to the very foundation of our understanding; to the distinction between life and matter, between past and future. It points to the non-dualistic foundation of what is and to the temporal nature of the self. Derrida, through *differance*, uncovers the system that makes all meaning possible; he shows us the sameness in difference that marks not only words and concepts, but also our understanding of reality.

Additionally, when we actively remember (use our working memory), just as when we are reading a text (also, using our working memory), we sort through competing codes in order to gain understanding. And every memory that makes up who we are must be reinforced in order to become part of the system, to become part of the self. The very act of looking, touching, smelling, hearing, tasting, experiencing, and remembering is an act of affirmation that attaches meaning to the stimuli or signs. And, as each memory is again and again affirmed, it becomes more of a part of us. Concepts and stimuli that are activated, felt, experienced together become connected. We then ascribe value and meaning and that too becomes connected to the concepts. “Old memories are the result of accumulations

of synaptic changes in the cortex as a result of multiple reinstatements of the memory” (LeDoux 107).  

In the same way, education forms memories which are connected to other memories and reinforced. Early in life we are taught, or molded, in school and in the home to form certain concepts and dualisms and to ascribe particular values to these memories. These accepted abstract concepts and dualisms (like justice and injustice, meaning and meaninglessness, mastery and submission, true and false) are passed down through the generations with ascribed values; indirect experience is learned through these direct experiences. Culture is passed down through language in the form of public narratives, such as myths and legends, and within the very words that we use. The privileging of one concept over the other, and the grouping of different concepts by one signifier, is a form of shaping and molding the minds of a people in a particular culture. We learn to privilege not only one concept over the other but also one type of discourse over the other. 

This does not mean that all concepts are relative or that they do not have value. I am simply showing the process by which value is formed in order to show the effects that education and culture have on the creation of the self. This shaping by society should be recognized because if we do not understand the process, then we will not be able to fully understand the development of the self. And also, we will not be able to effectively question the validity of the values that are shaping each of us and to actively change these values when they are harmful or reinforce these values when they are beneficial. As Foucault pointed out, we are taught what is acceptable and what is taboo, who is deemed sane and who is deemed insane, who can speak and who can not, and what is a valid area of inquiry and what is not.  

For Foucault the three ways that discourse is controlled include: prohibited words, the division of madness, and the will to truth (Foucault 339-342). As mentioned above, words have power because they can shape the minds of people and change the very pattern of your synapses. According to Foucault, words and discourse as a whole are controlled in order to avert that power. In The Discourse on Language, Foucault wrote, “I am supposing that in every society the production of discourse is at once controlled, selected, organized and redistributed according
to a certain number of procedures, whose role is to avert its powers and its dangers, to cope with chance events, to evade the ponderous, awesome materiality” (Foucault 340).

It is important to note that the creating of concepts and binary oppositions, the joining together of different concepts under one signifier, and the attaching of value and meaning are ways that the very development of the self is molded by the society and culture that each of us is a part of. It molds the very pieces, the words and experiences, that we use to create the narrative self and even implicit memories are formed within a system that is in many ways influenced by culture. These facts make clear the importance of fully understanding the concepts and assumptions behind the words that we use in communication because of their deep influence on the development of the self.

Therefore, we are molded by our culture, by the environment around us, by our DNA, and by ourselves because the self is an active creation of the self. We pick and choose the meanings that our explicit memories have. We actively create our narratives by ascribing meaning, by choosing (to some extent) which memories are reinforced and which are not, and by questioning and changing the values that we live our lives by. Memories are not reality itself but rather the impressions of reality made upon the mind; some of these impressions must be consciously affirmed in order to gain meaning, while others are unconscious impressions implicitly reinforced by experience. These impressions are not reality, just as words are not reality, but both reflect reality and are a way to understand and give meaning to both the self and the world.

The self, who you are, is reflected in the dynamic interplay between nature and nurture, between the semiotic and the symbolic, between the explicit and the implicit, between society and the self, and between the narrative and interpretation; each of us is unique, inseparable from the context of our lives, from our time and place, and from the system of differences that makes the creation of the self possible. Derrida quoted Heidegger as writing, “Being speaks through every language everywhere and always” (qtd In Derrida 464). 73 Being does speak through every language because language is a reflection of the system of the mind, of the self, and in that way it reflects being. It is at the same time a manifestation of the self, the narrative identity of a culture, and the system that is essential to the understanding of reality and to the creation of that self.

It is no wonder that the question, “Who are you?” cannot be answered with a few words. Words can only indirectly communicate the vast system that makes up the self, with pale impressions like shadows on cave walls. Language enables us to communicate our thoughts and at the same time denies us precise meaning. In that way, the unique tapestry that makes up the self can never be fully communicated to another. Each of us can never be known, truly and fully known, because we are ourselves signifiers within a complex system of society, objective beings whose rich and deep meaning is constantly deferred down an endless chain that is continually in the act of simultaneous creation and deconstruction. “Who are you?”, then, must be answered with a few words, because to communicate who we truly are would be an impossible task. So, when you are asked that seemingly simple question, you have no choice but to answer with your name, which is itself a pale signifier of the awesome and complex tapestry that makes up who you are.

References


Searle's Chinese Room Argument and its Replies: A Constructive Re-Warming and the Future of Artificial Intelligence

Mariano de Dompablo Cordio

Philosophy
University of Wisconsin, La Crosse

Abstract

This paper treats the philosophy of John R. Searle in his article "Minds, Brains, and Programs". It shows, using Searle's Chinese room argument (CR), that what Searle calls strong artificial intelligence (AI), the thesis that minds are to brains as computer software is to computer hardware, is not only false, but also that it must be false. The CR does this by arguing in effect that there can be no translation between Chinese and John's English understanding, and likewise neither can computers understand any question put to them because any question addressed to them is like the Chinese to John. On the other hand, ask John in English, which he understands as well as any native speaker, 'Do you understand Chinese?' and he will answer 'No.' What is the difference between John and computers? While John does not understand Chinese and does understand English, computers understand nothing. And because any programming language to computers is like Chinese to John, human-like behavior of a computer charged with running a robot, for example, proves nothing in the way of human understanding on the part of computers.

Because the CR does such a good job of proving the falsity of strong AI, a fundamentally different approach to the creation of AI is necessary. But, this is problematic for strong AI, namely, it leaves strong AI definitively behind.

1. Introduction

When we ask, 'Is artificial intelligence (AI) possible?' we really ask 'Can we create consciousness in computers?' This, as I see it, has been established by the philosophy of AI discussion so far. So, can we? And, if we can, what would we be
doing with the computer (if manipulating its inner structure or otherwise) in order for consciousness to be created with it? These questions put us in a place to define the AI project in two goals, a primary and a secondary. The secondary goal is a step toward the primary goal, consciousness in a computer. The secondary goal can be either of two possible alternatives as defined by the AI philosophy discussion so far: 1. writing a program, which when implemented in a hardware, results in a conscious computer according to test by observation OR 2. manipulating the inner structure of a computer in order to in effect reproduce within the computer the context for consciousness in a brain, providing the context necessary for consciousness but doing so in a computer. Both alternatives have an important background as they try for the primary goal. 1. takes consciousness (and a materialistic, biological account) for granted to a significant extent, and says that if behavior of the disputed conscious entity (the computer) can be mistaken for a human being given the proper circumstances, that entity understands (Turing 212). 1. says, 'Write the right program (one that programs for human-like behavior), and you have a conscious computer.' 2. assumes that consciousness is a property of the context provided by brains (and allows for a materialistic, biological account of consciousness).

In order to prevent any fudging on the definition of what strong AI (the thesis that minds are to brains as computer software is to computer hardware) supporters have as the goal of their projects, what I defined above as the primary goal, it is important to understand that strong AI projects and strong AI itself seem to be rooted in large part in science fiction. This happens in two ways. First, in so far as strong AI seems to borrow ideas for its projects like that of consciousness in computers from science fiction, it is innocuous. Ideas from science fiction examples of AI are not all that is borrowed. Second and more important is the very idea, the computer model of the mind, that hardware performing computation according to software-indicated directions could be intelligence, which may also be a holdover of too much science fiction, is damaging to a more biological account. But a deeper investigation into the origin of models of the mind over the course of history will have to be left for another time. To better pin down the goal of the projects of computer science from another direction, refer to the technical language used to name and describe computer scientists' AI projects. Some examples are: genetic programming, memory, master/slave programs, teaching programs, and programs that learn. All of these examples exude something of a human sense, a personification of their projects. In these examples, I have assumed that we see what computer scientists believe their projects can or will be able to do. But there is a problem with their usage of that language.

What computer scientists are trying to do with their projects is evident,
realize the primary goal, and that is illustrated by their language. The problem seems to be that computer scientists do not understand the significance of the language that they use, which is telling of an incomplete picture of the mental properties involved in their language e.g. consciousness or qualitative, subjective states of awareness and intentional states. So, to make a proper attempt at realizing the mind or its property consciousness in their computer would be almost happenstance because they do not understand the goals implicit to their language when, for example, they describe programs as teaching or learning. This taking for granted of one of these goals, consciousness, in the attempt to create AI reaches yet farther into the interdisciplinary field, cognitive science, of which computer science is a part. The philosopher branch of cognitive science as well often leaves out or inadequately treats in its philosophy of mind the What is it like? or the qualitative experience had by consciousness (Nagel 321). Conscious understanding and subjective intentional states are not, however, left out by the philosophy of John R. Searle in "Minds, brains, and programs" (235). His wonderful article is where this paper will begin.

2. Chinese Room Argument

Before continuing to my adapted rendering of the Chinese room argument appearing in Searle's article, the reader should understand that the Chinese room that Searle describes in his argument is designed to be identical in principle to any computer. Thus, anything that the Chinese room can or cannot do parallels all relevant computer capacities. A person, John, is in a room. John does not understand Chinese symbols nor is he capable of recognizing Chinese symbols such that he can distinguish Chinese symbols from Japanese symbols nor is there anything "To [keep John from believing that] Chinese writing is just so many meaningless squiggles" (Searle 236). Also in the room are two windows and two boxes. In Box 1 are directions written in English and divided into sets. Each particular set of English directions correspond to an attached Chinese symbol also inside Box 1. In Window 1 comes a Chinese symbol which John receives. John proceeds to Box 1, matches the Chinese symbol received to the same type Chinese symbol to which are attached a particular set of English directions. John follows these directions, which he understands as well "as any other native speaker of English", by finding in Box 2 the direction-indicated symbol (Searle 236). John then proceeds to Window 2 and puts this symbol out the window so that John effectively correlates one set of formal symbols with another set of formal symbols, and all that 'formal' means here is that [John] can identify the symbols entirely by their shapes . . . [answering] by manipulating uninterpreted
formal symbols. (Searle 237)

And, all of this constitutes John simply behaving as a computer, performing "computational operations on formally specified elements" (Searle 237). So, John successfully performs computation without understanding Chinese. After all, his responses to this point have been indicated by the English directions. But what if we gave John questions written in English?

Suppose that we ask in English 'John, do you understand Chinese?' and indicate that he should pass his answer out Window 2. John would answer 'No.' What is the difference between John's response to the question in English and his response to the same question instead addressed in Chinese? Ask yourself what it would be like to be John, comparing both experiences; John experiences understanding of the English question and his answer. He understands nothing of the same question addressed in Chinese. Nor does he understand his Chinese answer as anything more than the pushing of a shape out of a window that reads "Window 2" over it. What does Searle provide us in this thought experiment? A distinction between human understanding of a familiar language versus a language one does not speak. It is also important to notice that computers and humans share an engagement with the shape of written words, but unlike John who understands the English and not the Chinese, computers cannot understand any language even though they work in them. These considerations provide us with an appropriate setting for a few of the many replies to the Chinese room argument.

3. Objections to the Chinese room argument

The first of the objections to the Chinese room argument to be treated will be the systems reply as named by Searle. The systems reply concedes that the man inside the Chinese room does not understand the Chinese version of the question put to him 'Do you understand Chinese?' However, John is but a part of the system, and the whole system, the Chinese room and any digital computer by way of its having at its disposal everything accessible to a digital computer, understands the question.

Searle replies, let John memorize every system element, the English directions and attached Chinese symbols in Box 1 and the Chinese symbols in Box 2 so that the composite's aspects comprise all aspects of the entire system. John still does not understand Chinese. There simply can be no translation between John's English understanding and the Chinese symbols, no matter John's Turing test mistaken understanding. And, "a fortiori neither does the system [understand], because" anything in the system is part of John (Searle 240). So, strong AI is false, the systems reply fails, and the Turing test has counterexamples, so it is
ineffective. With regard to the systems reply, Searle makes some additional remarks, which I think outline what has been the state of these affairs.

It is not easy for me to imagine how someone who was not in the grip of an ideology would find the idea at all plausible. Still, I think many people who are committed to the ideology of strong AI will in the end be inclined to say something very much like this; so let us pursue it a bit further. (240)

The systems reply replies: "'the man as a formal symbol manipulation system' really does understand Chinese." (Searle 240) In this reply, the systems reply begs the question, that is, it insists the truth of its claims without argumentation in addition to its original argument. So, the systems reply is false. There are additional comments made by Searle with regard to objections to the Chinese room that do a great job of outlining the central errors of strong AI supporters.

The robot reply is the second objection to the Chinese room argument. It asks us to think about a new program. This computer, with its program written not only for the taking in and putting out of symbols, would perform the function of operator of the robot in which it is placed. The computer would operate the robot in such a way that its behaviors are similar and can be confused for something with human-level understanding. The idea is that this computer instead of the original digital computer treated by the Chinese room argument would have understanding. What should be noticed about the robot reply?

The robot reply indirectly claims that cognition is about being-in-the-world with certain realities of what it is to be a causal force in the world instead of just formal symbol manipulation (Searle 243). The obvious reply to the robot reply is that putting another computer inside of a robot does not get rid of the original problems outlined by the Chinese room argument; innovation in programming whatever cannot improve upon the problems which have thus far been outlined and are essential to all algorithm based approaches to the creation of AI, strong AI. In order to illustrate this criticism, imagine that inside that robot is John instead of the extra computer with its new program. The presence of John in the room for the purpose of carrying out the computation needed for the robot's operation is in principle equivalent to the needed computation otherwise being carried out by the computer. After having replaced the computer with John, follow through with the original Chinese room argument, that is, to the conclusion that there can be no translation between John's English understanding and Chinese, and one will understand that the robot reply is false. The implicit notion underlying the robot reply (as in the systems reply) that 'If it behaves like it, it must be it
(understanding) does nothing to improve upon the essential state of John in the room and strong AI by the same force of the Chinese room argument. Given strong AI's falsity, an adjusted approach to the creation of AI will be necessary since strong AI must be false.

4. Further inquiry into the possibility of AI

A fundamentally different approach to the task of creating AI may be that of manipulating the orientation of the firings in the hardware of a computer in order to reproduce the necessary physical context for consciousness in brains but doing so in a computer. Though this approach may not reproduce the physical context necessary for consciousness because it would only reproduce the electrical portion of what would presumably be an electro-chemical reality in the brain. However, the electrical approach seems plausible in our pursuit of the reproduction of conscious subjective states necessary in order to in some way produce the understanding necessary for what is the precise primary goal of AI, intelligence. But this approach is problematic for strong AI, namely, it leaves strong AI definitively behind. Strong AI takes it as its assumption that hardware can realize certain desired properties of the brain without a neurophysiological account of the brain to reach the same goal. This adjusted approach leaves behind the use of just computer hardware in order to provide a physical context for consciousness, which is not just computer hardware proper. Without consciousness, however, intentional states would be had by nothing. So, while AI seems plausible with the appropriate context for consciousness, neurophysiology is required before it can be created. Strong AI cannot be true.

5. Conclusion

The original assumption of strong AI is that certain desired mental properties can be achieved using hardware in a computer instead of a brain, that brains are effectively hardware. But, this view, strong AI, has been proven false on multiple occasions by John R. Searle's Chinese room argument. The Chinese room argument has shown that there can be no translation between John's English understanding and Chinese, making understanding impossible at either John's level, that of the central processing unit (CPU), or at the level of the entire system, that of the CPU, the wires, and any additional hardware used in computation. Likewise, understanding is impossible when placing a computer inside of a robot so that it may move around in the world, for if John were in the room performing the same computational tasks, he could not understand Chinese, and moving around in the world in a human-like fashion does not improve upon the state of computers. The falsity of this last example, the robot reply, and the notion that mistaken behavior is a sure indicator of understanding casts serious doubt on the
Turing test. So, an adjusted approach is necessary for the successful creation of AI. The approach described in this paper, however, is a fundamentally different approach to that of creating strong AI, as manipulating the orientation of the firings in the hardware of a computer leaves the original assumption of strong AI behind, thus leaving strong AI definitively behind. Indeed, the re-orientation of firings in hardware on the model of a neurophysiologically defined orientation of neuron firings in order to provide the appropriate context for consciousness while using a materialistic, biological account of the brain is not hardware proper as required by the strong AI thesis. Rather, it is a synthesis of hardware and neurophysiology. This approach may not even reproduce the necessary physical context for consciousness because it only reproduces the electrical portion of what would presumably be an electro-chemical reality in the brain. So, AI (and not strong AI) is possible in principle, but it is dependent on a materialistic, biological account of the physical context for property consciousness as defined by a consciousness as a property of the brain hypothesis of the brain-consciousness relation. AI on the basis of the non-biological account of the brain in the strong AI thesis cannot be possible.

References

Thinking About Your Foolishness: Foolishness in Terms of Metacognition

Jelena Markovic

Philosophy and Psychology
University of Toronto

As long as there is self-awareness, there will be foolishness. With the ability to reflect on one’s existence comes responsibility and a normative standard of behaviour. One perhaps distinctly human dimension of self-awareness is metacognition. ‘Thinking about your thinking’ is involved in regulating complex patterns of behaviour which may be maladaptive, self-destructive, or a poor expression of ability. In other words, metacognition is involved in regulating behaviour that is foolish. It also has the power to reveal potentially foolish thoughts that are incorporated in these patterns. So even at first glance metacognition and foolishness are related. To understand how they intersect we first need to explore metacognition in greater depth. However, the idea pervading this exploration is that foolishness can be viewed in terms of deficiencies in the metacognitive process. We can begin by first better defining foolishness, and then applying relevant metacognitive models to this definition.

Foolishness, to be used as a meaningful psychological concept, must fulfill two requirements. The first is that it be manifested in a pattern of behaviour. There are of course, foolish acts, but it is not productive to talk about the cognitive mechanisms operating behind them, as an isolated foolish act can frequently be explained away by irregularities of the situation. The second requirement is awareness. The person who behaves foolishly must have the knowledge that what they are doing is foolish; otherwise the behaviour can be attributed to ignorance, which is a very different problem. The individual’s awareness incorporates motivational factors that keep the foolishness mechanism running despite knowledge that the behaviour is maladaptive. This character is the essence of foolishness as it is used in psychological literature. The two factors defined here,

2 Perkins (2002), for example, makes the distinction between blind folly and plain folly.
knowledge of foolishness and its persistence can be readily examined in terms of metacognition.

Fernandez-Duque et al., (2000) divide metacognition broadly into metacognitive knowledge and metacognitive regulation. Foolishness as a persistent pattern of behaviour can be considered in terms of metacognition as a regulatory process, and the awareness of the foolishness of one’s behaviour can be looked at as metacognitive knowledge.

To examine metacognition in terms of regulation or controlled processing, it can most generally be divided into two components: metacognitive monitoring and metacognitive control (Nelson, 1999). Nelson and Narens model metacognition hierarchically in terms of an object- and meta-level of awareness (Nelson & Narens, 1994); monitoring and control are the mechanisms by which the two levels interact. Metacognitive monitoring encompasses evaluative judgments of one’s cognition and metacognitive control is the implementation of these judgments. Nelson, among others, takes a functional approach to metacognition, meaning that monitoring is thought to precede control. The monitoring process examines and analyses cognition, then affects it through the control process. This model can be used as a framework for looking at foolishness, in particular, the way foolishness is manifested as a self-regulation failure.

Various studies examine the relationship between brain areas responsible for meta- and higher cognition and self-regulation. An interesting case is Beauregard’s 2007 study on the effects of metacognitive strategies on arousal during exposure to erotic film clips. FMRI imaging was used to see which brain areas were activated during the use of metacognitive “detachment” strategies to suppress arousal in contrast with a control group that experienced the stimuli normally. The use of metacognitive strategies was linked with increased activation of the anterior cingulate cortex and right lateral prefrontal cortex. The prefrontal cortex is implicated in several top-down control processes, most importantly in the inhibition of inherent response tendency and emotional regulation (Davidson et al., 2000; Damasio, 1995). The target group, unlike the control group, showed no significant activation in the amygdala (an emotion processing area). Thus, Beauregard has demonstrated that metacognitive strategies change the way arousing stimuli is experienced in the brain and can activate brain areas responsible for behavioural control.

At first glance it may seem as if we have successfully applied Nelson and Naren’s model of metacognition to view foolishness as a failure in self-regulation: The foolish individual receives monitoring information and fails to use proper control strategies to regulate their cognitions. This leads to
foolish thoughts and thus foolish behaviour. Was that not the idea? The problem with this account is that painting foolishness as a failure in self-regulation quickly emerges as overly simplistic. After all, self-regulation can be foolish as well. One can’t deny that an obsessive dieter who is willing to skip meals and try hazardous drugs is self-regulating. One can call him foolish, however. The persistent, self-organized brand of foolishness defined above cannot be productively examined in terms of a malfunction in self-regulation. ‘Malfunction’ is a poor word choice since the process itself is functioning just fine. But how can that be?

The claim that foolishness has nothing to do with a failure in self-regulation may not seem plausible under the traditional account of self-regulation, namely, that it is a proper assertion of willpower. This is a standard notion in folk-psychology and psychology alike. For instance, Ayduk and Mischel (2002) ascribed foolishness to situations in which one’s willpower fails to reign in the ‘hot’ impulsive, emotional system despite the fact that the individual knows the right action to take. This account may seem plausible, but upon further investigation, it does little to explain the kind of foolishness outlined earlier. If foolishness can be remedied by use of one’s willpower, why do individuals continue to display maladaptive behaviour? What mechanism explains the foolishness itself, as a self-organizing pattern? These questions are better addressed to a more sophisticated model of self-regulation.

Recent neurological research (Lewis & Todd, 2007) views self-regulation as a coordination among brain systems. Successful self-regulation is defined in terms of where the epicenter of control is for the self-regulation process rather than a ‘hot’ emotional system subservient to a ‘cool’ cognitive one. In this case, foolish behaviour is a result of the self-regulation process being centered in limbic areas and the bottom-up regulation of cortical activity of inputs from the amygdala. This is what is referred to in folk psychology as “weakness of will” or “overwhelming emotions” (discussed in Perkins, 2002). The classic way in which we view self-regulation (i.e. domination of the ‘cool’ system) is as coordinated brain activity for which the epicenter of control is the anterior cingulate cortex and other areas of the prefrontal cortex. These brain regions allow for purposeful control of one’s actions and thus the top-down regulation of limbic (emotional) processes (Lewis & Todd, 2007). Thus, through this model, foolishness is a self-sustaining pattern of behaviour because it is a self-regulation mechanism rooted in the ‘wrong’ part of the brain.

This model provides a more detailed account of how the brain areas that we associate with cognitive control really function in a more integrated
manner than we first imagined, however it remains incomplete. An individual can have a self-regulation process whose locus of control is in the prefrontal cortex and which is characterized by top-down control of limbic areas but she can still display foolish behaviour if the cognitions at the top are wrong. Foolish self-regulation is procedurally no different from wise self-regulation. It is the specific beliefs and emotions involved that make the difference. This is why it is crucial to look at how one’s metacognitive knowledge interacts with the control/monitoring system.

Dweck provides an excellent illustration of the dramatic effects of one’s metacognitive knowledge on self-regulated behavioural patterns. Commenting on the self-destructive beliefs that “make smart people dumb”, Dweck cites poorly informed metacognitive beliefs as a source of foolishness in intelligent individuals (2002). More specifically, a belief that intelligence is fixed and that performance evaluations measure overall intelligence instill an aversion to situations that may expose intellectual inadequacy. Thus, this item of metacognitive “knowledge” instantiates a behavioral pattern of challenge avoidance that results in eventual atrophy of one’s cognitive abilities. The remedy is a metacognitive belief that focuses on effort and the plasticity of intelligence, and thus makes the individual more likely to place themselves in situations that challenge and expand their mental abilities.

Dweck’s studies, although encouraging, spur questions. How did the ‘foolish’ participants come across their metacognitive beliefs? And how did those beliefs act to instantiate a destructive pattern of behaviour? The latter question is an important one since foolishness, as mentioned earlier, requires awareness. It is a case of having the requisite metacognitive knowledge where that knowledge is not reflected in one’s behaviour.

Perhaps the questions above are not for Dweck, but rather for Nelson and Narens. The functional account of metacognition does not explain how metacognitive knowledge relates to the control/monitoring process, nor does it match up to more holistic views of self-regulation emerging in neuroscience. Recently, researchers have acknowledged that metacognitive knowledge and related awareness seem to be crucial to the function of the metacognitive regulation system (Fernandez-Duque et al., 2000, Panaoura & Philippou, 2007). Another problem for the original model is that it is hierarchical and ordered, stating that metacognitive monitoring processes influence control processes (which is how metacognition influences cognition). Koriat et al. (2006) have tried to bring awareness to the reverse process, by which control processes influence monitoring. They call this process CM, as opposed to the traditional MC model. Furthermore, they
delineate two ways in which metacognition can function: the sequential model, in which metacognition reverts from MC to CM in sequence, and the simultaneous model, in which both processes function at the same time. Koriat et al. present a much-needed look outside the traditional boundaries of how metacognition functions but they have just begun to scratch the surface. The sequential model is still a hierarchical one, and no definitive account of the simultaneous model is provided. Thus no parallel model for metacognition has definitively emerged. The importance of a parallel model will be further clarified in the context of how a self-destructive pattern is defeated, but it has not gone unnoticed in scientific literature. Dent (2003), for example, has urged social scientists to consider the inadequacy of a linear cause-effect framework in explaining complex phenomena and to instead adopt a mutual causality model.

An example of a parallel dual-process model is the one used by Stanovich and West (2000) in explaining the gap between descriptive and normative rationality present in individuals. Two processes function in parallel: an efficient, heuristic-based System 1 and a flexible, analytic System 2. Systematic individual differences in rationality are a result of differences in coordinating the two systems at the intentional level as opposed to differences in intelligence present at the algorithmic level. In a similar vein, we can imagine a parallel dual-process foolishness model in which intentional-level cognitions and metacognitions monitor and control each other simultaneously while both maintain a balance between System 1 and System 2 processes.

This more complex model is well suited to fill the gaps left by the original. The effect, or lack thereof, of metacognitive knowledge is clarified by a distinction made by Teasdale (1999) between metacognitive knowledge and metacognitive insight. Metacognitive knowledge is present only in a declarative form; it can have no real effect on the controlled processing of the organism. In order for a real change to occur, i.e. in order for the metacognitive knowledge to be manifested, one needs to have metacognitive insight at the intentional level. Only insight allows for the knowledge to be put into action and this insight can only be gained through direct experience. This effectively addresses how it is that foolish behaviour persists in the presence of metacognitive knowledge. It is due to a lack of insight. We have seen the complexity of a self-regulation system and know that metacognitive knowledge is not enough to disrupt such a system already in place. That disruption would require an equally complex system, one that can only be built within a parallel model. This is because insight arises from a shift in thinking mode and not only an improvement of the self-regulation process.
For if foolishness were merely a matter of self-regulating behaviour processes, this model would solve nothing. One would have to set up successful self-regulating systems for all aspects of life vulnerable to foolishness, which would be an overwhelming if not absurd task. What is needed to overcome foolishness is a meta-level of regulation; and this is exactly what a dynamic model provides. The individual is no longer constrained to situation-specific instances of insight because insight is a case of not only different subject matter for thought but a different way of thinking. Metacognitive processes functioning in parallel with cognitions allow a person to see the effect of her current cognitions. And this is the level of awareness required to circumvent foolish behaviour. It is distinct from the awareness present in having metacognitions or knowledge that one is acting foolishly since it encompasses the entire regulatory process. Awareness of the effect of cognitive processes allows one to unravel them if they are destructive. Otherwise, lack of meta-awareness permits foolish patterns to perpetuate.

The notion of insight as the key to banishing foolishness has a rich heritage. Awareness of one’s thinking is at the core of the Buddhist meditative tradition, which is why Teasdale refers to a mental process imbued with metacognitive insight as mindfulness. Another comparison that can be drawn is that of process-oriented thinking, in which one is no longer focused on the output of cognition (product-oriented) but rather on the process of thinking itself. Process-oriented thinking is responsible for Dweck’s participants’ beneficial metacognitive beliefs, since their focus is the process of learning and not the appearance of intelligence. Finally, this form of insight goes back to Plato, whose Socratic dialogues were aimed at dispelling foolish beliefs in both Socrates’ conversation partner and Plato’s audience. Many have wondered why Plato ended his dialogues with *aporia* as opposed to providing an account of piety, say. What is accomplished in leaving the reader in the cold? The answer is simply: Even if Plato were to tell you what piety is, you would have no reason to believe him. In order to have a *logos* of piety, you need to reflect and experience it for yourself. This is the nature of insight and the reason why without it, foolishness persists.

References


Indiana Undergraduate Journal of Cognitive Science

2007 – 2008 Editorial Board

Executive Editor
Brenden Sewell, Indiana University

Student Reviewers
Sanjeev Murugan Balamohan, Indiana University
Jordan Thevenow-Harrison, Indiana University
Patrick Mundy, Indiana University

Faculty Sponsors
Dr. Ruth Eberle, Director of Computer Technology and Adjunct Professor of Cognitive Science & Informatics, Indiana University
Dr. Robert Goldstone, Chancellor’s Professor and Director of Cognitive Science Program, Indiana University

Web Design / Programming Support
Michelle Capriles-Escobedo, Web Designer, Indiana University
Muhammad Nubli Mohd Kasa, Systems Analyst/Programmer, Cognitive Science Program, Indiana University
Indiana Undergraduate Journal of Cognitive Science
http://www.cogs.indiana.edu/iacs/journal.html

Author / Submission Instructions

I. General Information

The Indiana Undergraduate Journal of Cognitive Science invites submissions of original writing by undergraduate students. Submissions may come from any area within Cognitive Science including, but not limited to: artificial intelligence, anthropology, biology, computer science, linguistics, philosophy, psychology and neuroscience.

II. Submission / Paper Format

Articles are accepted on a continuous basis and will be considered for publication upon submission. Articles should be sent directly to the editorial board as an attachment in Microsoft Word or Adobe PDF format. Submissions should be edited for grammar and style before submission. There is no limit on article length. Submissions should include a Title Page that includes the following information: Article Title, Author Name, Major, University, and E-Mail Address. This information will not be published and is for contact purposes only.

Authors should submit their work via E-mail to the Indiana Undergraduate Journal of Cognitive Science Editorial Board at iacs@indiana.edu. Once your submission is received, a confirmation E-mail will be sent by the Editorial Board. All submissions will be considered equally and no preference will be given to any particular discipline within cognitive science.

III. Review / Acceptance Process

After submission, the Editorial Board and a panel of reviewers will review all articles and will decide which papers will be published in the current edition of the journal. Authors will be notified by E-mail if their paper is accepted for publication. Authors will also be notified by E-mail if their paper is not selected for publication. After acceptance, authors are required to submit their paper in Microsoft Word format to allow the Editorial Board to make formatting and grammatical edits. Once these changes have been made, the Editorial Board will contact the authors to obtain final approval of the above changes and to obtain written publication permission.

IV. Disclaimer and More Information

Articles published in the Indiana Undergraduate Journal of Cognitive Science are considered copyrighted. However, this journal is not a binding publication. Authors are free to submit their work to any other publications they wish.

For more information about the journal, please contact Brenden Sewell, Executive Editor of the Indiana Undergraduate Journal of Cognitive Science, at brrsewel@indiana.edu or by the means below.

Indiana Undergraduate Journal of Cognitive Science
819 Eigenmann Hall - 1910 E. 10th St.
Indiana University - Bloomington, Indiana 47406
E-Mail: iacs@indiana.edu http://www.cogs.indiana.edu/iacs