Philip Woodward[1]

Consciousness and Rationality

The Lesson from Artificial Intelligence

Abstract: *I review three problems that have historically motivated pessimism about artificial intelligence: (1) the problem of consciousness, according to which artificial systems lack the conscious oversight that characterizes intelligent agents; (2) the problem of global relevance, according to which artificial systems cannot solve fully general theoretical and practical problems; (3) the problem of semantic irrelevance, according to which artificial systems cannot be guided by semantic comprehension. I connect the dots between all three problems by drawing attention to non-syntactic inferences — inferences that are made on the basis of insight into the rational relationships among thought-contents. Consciousness alone affords such insight, I argue, and such insight alone confers positive epistemic status on the execution of these inferences. Only when artificial systems can execute inferences that are rationally guided by phenomenally conscious states will such systems count as intelligent in a literal sense.*

Correspondence:
Email: pwoodward@niagara.edu

[1] Department of Philosophy, Niagara University, USA.

Journal of Consciousness Studies, **29**, No. 5–6, 2022, pp. 150–75
DOI: 10.53765/20512201.29.5.150

1. Introduction: The Problem of Consciousness for AI

The digital age has been marked by an ongoing debate over whether the term ‘intelligence’ could literally apply to machines. Optimists say yes: any system with cognitive capacities over a certain threshold, natural or artificial, counts as a form of intelligent life, and machines will surmount that threshold at some point (if they have not already). Pessimists say no: however well our machines *simulate* intelligence, they do not, and perhaps could not possibly, meet all the necessary conditions for literal intelligence.

Perhaps the oldest and most persistent source of pessimism is the intuition that consciousness is necessary for intelligence, but that artificial systems lack consciousness. Alan Turing mentions an early version of this source of pessimism in his famous 1950 paper ‘Computing Machinery and Intelligence’. He quotes Jefferson Lister from the previous year:

> Not until a machine can write a sonnet or compose a concerto because of thoughts and emotions felt, and not by the chance fall of symbols, could we agree that machine equals brain — that is, not only write it but know that it had written it. No mechanism could feel (and not merely artificially signal, an easy contrivance) pleasure at its success, grief when its valves fuse, be warmed by flattery, be made miserable by its mistakes, be charmed by sex, be angry or depressed when it cannot get what it wants. (Turing, 1950/2010, p. 338)

Call Lister’s worry ‘the problem of consciousness’. Turing dismisses Lister’s worry on the grounds that if it were on point, we would be left without an empirical criterion for attributing intelligence. This consequence, if it is indeed a consequence of Lister’s position, would be unfortunate, but would hardly amount to a refutation. But there are plenty of other concerns one might have about Lister’s remarks. Exactly which sorts of conscious states are important for intelligence? What functional role do they need to have in the system? If conscious states did play this role in a machine, how would we know? And most importantly, even if consciousness is central to *human* intelligence, why think it is essential to intelligence as such? If the problem of consciousness is a real problem, we would need to get much clearer on the details.

Arguments that appeal to the problem of consciousness flow in a particular direction: from a claim about what full-fledged intelligence involves, to a conclusion about the limitations of actual or potential machines. Proponents of such arguments tend to pay only as much attention to the details of how computers function as they must (which does not go unnoticed by the AI researchers whose life work is getting criticized). A very different kind of argumentative strategy in support of pessimism flows in the opposite direction: from the details of how computers function, to the conclusion that such functioning does not or could not scale up to full-fledged intelligence. Here the technical details matter. The two most influential pessimistic arguments have proceeded in this direction. The first is 'the problem of global relevance', pressed by Hubert Dreyfus and Jerry Fodor. According to Dreyfus and Fodor, given the nature of computation, it is impossible for a machine to be able to solve problems that require determining, from an unbounded pool of information, which bits of information are relevant to the problem's solution and which are not. The second is 'the problem of semantic irrelevance', pressed by John Searle. According to Searle, given the nature of computation, it is impossible for a machine to understand what its symbols mean. But — as both of these arguments proceed — these impossibilities for machines correspond to the heart and soul of intelligence. Both arguments have been subjected to vigorous critique from AI researchers, as we will see.

My task in what follows is to see what happens when these two argumentative strategies — the reverse-engineering strategy based on the problem of consciousness, and the forward-engineering strategy based on the problems of global relevance and semantic irrelevance — are lain on top of each other, so to speak. I suggest that the best way to develop the problem of consciousness is to understand the cognitive role of consciousness along the lines of what, according to the other two arguments, is missing from artificial systems, and vice versa. This way of connecting the dots among these three pessimistic arguments reveals, I will argue, an important connection between consciousness and intelligence: consciousness alone makes it possible to grasp rational connections among representational contents, and such grasping plays an ineliminable epistemic role in any genuinely intelligent system. This point, if correct, places a serious constraint on the artificial intelligence enterprise. But it is also of independent interest to philosophy and psychology.

Two points of clarification are in order before we proceed. First, the term 'intelligence' is used in various ways. One basic distinction is between two ways of thinking of the standards by which the intelligence of a system is measured: a consequentialist standard, in terms of whether the system gets the right results, and a deontological standard, in terms of whether the system gets its results in the right way. My focus will be on the latter. The term ‘rationality’ is standardly used to denote intelligence in this deontological sense (hence my title).

Second, the term ‘consciousness’ is also used in various ways. It is standard to distinguish between ‘access consciousness’ and ‘phenomenal consciousness’. The former is a cognitive property: a mental state’s availability for verbal report and other cognitive tasks. The latter is an experiential property: a mental state’s qualitative feel, what it is like for the experiencing subject to be in that state. I will, following Lister — who emphasizes ‘thoughts and emotions *felt*’ — focus on phenomenal consciousness.

Here is the plan. In Section 2, I introduce the problem of global relevance and describe the current state of the dialectic surrounding it. In Section 3, I discuss one putative solution to the problem of global relevance in terms of consciousness, *viz*. the global workspace architecture. In Section 4, I introduce the dialectic surrounding the problem of semantic irrelevance, including Searle’s suggestion that consciousness is what renders semantic properties cognitively relevant. In Sections 5 and 6, I propose my own view of the way that the three problems are related, and I defend this proposal against objections. In the concluding section, I summarize the implications of my discussion for the dispute between optimists and pessimists, and for psychology more generally.

2. The Problem of Global Relevance

Hubert Dreyfus first published his pessimistic arguments in ‘Alchemy and Artificial Intelligence’ (1965), and then developed them further in *What Computers Can’t Do* (1979 and subsequent editions). His worries cluster around the role that common-sense knowledge plays in human cognition. Here is one of his favourite examples (which he borrows from Douglas Lemat). Take the two sentences:

> Mary saw a dog in the window. She wanted it.
> Mary saw a dog in the window. She smashed it.

To any competent human speaker of English, it is perfectly obvious what ‘it’ refers to in each case (the dog, then the window). But now suppose that we were tasked with programming a machine to make the right guesses here. How would we do it? The rules of grammar do not themselves settle the matter. Rather, there is a rich set of background beliefs about dogs, people, and windows that provide the requisite context. Moreover, these background beliefs are relevant to how the input sentence is comprehended, not just to which inferences it generates once it is comprehended. Thus, for an artificial intelligence to be able to reliably parse input sentences in natural language, it must (a) have access to all potentially relevant background beliefs, and (b) be able to select from among them those that are relevant to the parsing of the input sentence.

Similar points can be made about the role of common-sense knowledge in practical deliberation. Here is an example from Tim Crane that illustrates the point. A bus-driver in Arizona was disciplined for not departing from his bus route when a passenger had a heart-attack on his bus, when in fact he was following the rule, set down by his employer, not to depart from his route without permission. Clearly, Crane thinks, the bus driver was operating with too strict a rule in mind, and any intelligent being — natural or artificial — must be flexible enough to depart from such a rule in certain circumstances. But how can the right degree of flexibility be encoded in an algorithm? Crane writes:

> It is absurd to suppose that the bus company should present the driver with a rule like 'Only leave your route if you have permission, unless a medical emergency occurs on board, in which case you should drive to the nearest hospital, unless the hospital is under siege from international terrorists, or unless there is a doctor on board, or … in which case you should …' — we don't even know how to fill in the dots. How can we get a rule that is *specific* enough to give the person following it precise directions about what to do… but *general* enough to apply to all eventualities? (Crane, 2015, p. 83)

We readily know how to behave in novel situations, without recourse to norms specific enough to cover all the details of the situation. The problem for AI is to codify these common-sense ways of comprehending and navigating the world in terms of *rules*, and Dreyfus doubts that this can be done.

Similar concerns have been raised by Jerry Fodor in *The Modularity of Mind* (1983) and more pointedly in *The Mind Doesn't Work That Way* (2000). Even though Fodor was one of the most sophisticated advocates of the computational theory of mind, he also raised doubts that that theory can be applied beyond *modular* mental processes — that is, processes restricted to a certain type of information, such as visual or grammatical information — to fully general rational processes. His arguments focus on the challenge of implementing *abductive* reasoning in an artificial system:

> Because of the context-sensitivity of many parameters of quotidian abductive inferences, there is typically no way to delimit a priori the considerations that may be relevant to assessing them. In fact, there's a familiar dilemma: Reliable abduction may require, in the limit, that the whole background of epistemic commitments be somehow brought to bear in planning and belief fixation. But feasible abduction requires, in practice, that not more than a small subset of even the relevant background beliefs is actually consulted. (Fodor, 2000, p. 37)

Dreyfus's pessimism about our ability to codify common sense in terms of general rules, on the one hand, and Fodor's pessimism about our ability to build a program that can determine what information is relevant to the problem at hand, on the other, are closely related. There are types of rational processes to which *any* of the subject's beliefs could be relevant. Such processes include abduction (a type of theoretical rationality), and deliberation about high-level ends and means (a type of practical rationality). In order to engage in such rational processes, the agent first (logically, if not temporally) must decide which of its beliefs to consider. But there doesn't appear to be any way to implement this task as a computationally tractable algorithm. One could try to implement it either (1) by locating it 'in the program', i.e. by listing a bunch of rules in the program that specify which bits of information are relevant to which rational processes, or (2) by locating it 'in the data', i.e. requiring that the program search all of its memory for information 'tagged' in a certain way. Contra the first option, Dreyfus argues that common sense cannot be codified as rules; contra the second option, Fodor argues that such a search-process would be computationally unfeasible. This is the problem of global relevance.

Dreyfus and Fodor are both targeting *classical* computational architecture. It is now a commonplace that connectionist architectures dispense with explicit algorithms. Instead, they generate correct output-judgments as the result of training, not via rule-following. And it's true that today's 'DCNNs' (deep convolutional neural networks), the sort that power internet search engines, are rapidly expanding the power of computers to accomplish apparently intelligent tasks — navigating automobile traffic, for instance. Notably, Dreyfus considers a move away from classical architecture as a vindication of his arguments rather than an objection to them. Here in the heyday of neural network machine learning, isn't the problem of global relevance 'old hat'?

Not quite. To date, the problem-solving power of DCNNs has only been demonstrated in the contexts of fairly narrowly proscribed tasks (determine whether the picture is of a face; predict the next line of text the user will type; etc.). DCNNs have yet not been developed that can implement fully general abductive and deliberative processes. Perhaps they will. But no engineer has yet shown that a connectionist system, as such, can solve the problem of global relevance. As far as this problem is concerned, optimists and pessimists are thus at a dialectical stand-off.

3. The Global Workspace Theory of Consciousness

Why think that the problem of global relevance is related to the problem of consciousness? A preliminary answer is that a popular empirical theory of consciousness seems to solve it.

Empirical psychology has made several fascinating discoveries about the cognitive role of consciousness in humans (and similar animals). It turns out that humans can perform the following tasks subconsciously (below the level of awareness) or non-consciously (while asleep or anaesthetized): reach for an object with the correct grip aperture (even when consciously misjudging the object's dimensions); perform habitual or expert actions (such as signalling a lane-change, returning a tennis volley); drive; untie knots; write (somewhat garbled) emails; read or hear familiar words and phrases and be influenced by their meaning. But there are tasks that humans cannot perform subconsciously or non-consciously, such as the following: learn a new skill; register an associative link between temporally separated events; follow a multi-step procedure; deliberate about rival action plans; pursue a long-term goal; read or hear *novel* sentences and be influenced by their meaning.

Over 30 years ago, in response to what we had discovered about conscious vs. unconscious processing in humans, Bernard Baars (1988) proposed that consciousness has the following functional profile:

(1) Sensory signals are highly integrated (i.e. cross-modally).
(2) Sensory signals are integrated with background knowledge.
(3) Consciousness has limited capacity.
(4) Novelty calls for conscious attention.
(5) Conscious information is available to many brain functions.

Baars theorized that consciousness exhibits this functional profile because it acts as a 'global workspace' for the mind. Lower-level, specialized processors pass their outputs along to a higher-level system that integrates them and broadcasts them back out to all the processors. Patricia Churchland provides the helpful metaphor of a meeting of department-heads at a corporation: each department-head quickly reports on her/his department's latest findings, and then they all return to work, with updated marching orders (Churchland, 2013, p. 242). The global workspace model has accumulated empirical support over the years, and (in broad strokes at least) enjoys widespread, if not unanimous, support within the empirical sciences of consciousness (see e.g. Dehaene and Naccache, 2001; Deco, Vidaurre and Kringelbach, 2021).

Baars has been quite explicit that his model has the resources to solve the problem of global relevance, because of the way it links serial and parallel processing. In a paper coauthored with Murray Shanahan, he writes that global workspace theory

> …explains how an informationally unencapsulated process can draw on just the information that is relevant to the ongoing situation without being swamped by irrelevant rubbish. This is achieved by distributing the responsibility for deciding relevance to the parallel specialists themselves. The resulting massive parallelism confers great computational advantage without compromising the serial flow of conscious thought, which corresponds to the sequential contents of the limited capacity global workspace. (Shanahan and Baars, 2005, p. 174)

Suppose Baars is right that (1) the global workspace model accurately captures the functional structure of consciousness, and (2) the global workspace model explains how relevance-determinations are computationally tractable. What follows?

On the one hand, the optimist will point out that the problem outlined by Dreyfus and Fodor has an engineering solution, at least in principle. If a system is organized in accordance with the global workspace model, all relevant information can be available to a central processor without any need for cumbersome search-algorithms or codified common-sense rules.

On the other hand, the advocate of the problem of consciousness will point out that this engineering solution invokes a model *of consciousness*. The problem of global relevance appears to reduce to the problem of consciousness. No consciousness, no global relevance; no global relevance, no genuine intelligence.

But this is too hasty, as the optimist will rightly point out. For it is not clear that the global workspace architecture, as such, implicates *phenomenal* consciousness. True, there is evidence of a correlation between phenomenal consciousness and the global broadcasting of neuronal information in the human brain (see e.g. Dehaene, Lau and Kouider, 2018). But that doesn't mean that the global workspace theory is itself a theory of phenomenal consciousness. Maybe it is a theory of access consciousness, with which phenomenal consciousness is roughly correlated. (Such is the explicit position of its leading contemporary defender, Stanislas Dehaene.) And even if any system with a global workspace architecture is *ipso facto* phenomenally conscious, it is the causal properties of the architecture, rather than its phenomenal properties, that make a difference *vis-à-vis* global relevance. Why couldn't a central (perhaps classical) artificial intelligence, hooked up in the right way to a whole bunch of specialized (perhaps connectionist) artificial intelligences running in parallel, solve the problem, whether or not the central processes are phenomenally conscious (as they happen to be in humans)?

4. The Problem of Semantic Irrelevance

In his famous 1980 paper 'Minds, Brains, and Programs', John Searle presents his 'Chinese Room' thought experiment. It goes as follows: Searle, who knows no Chinese, is imprisoned in a special room. He has been given a pile of 'output' Chinese symbols and a rule book (written in English) that tells him which of these symbols to output given certain 'input' symbols. Searle imagines that, with enough practice, he (or rather the whole functional system) could pass for a competent speaker of Chinese. But his Chinese outputs are not functionally based on any comprehension of what the symbols mean. And that, Searle contends, is exactly what is going on in a digital computer. Since semantic meanings *are* relevant to what genuine speakers of Chinese say, it follows that no amount of digital computing can ever mount up to human intelligence.

Like Dreyfus and Fodor, Searle is targeting classical computational architecture. But the thought experiment could easily be recast to target connectionist alternatives. In the alternate version, no rule book is supplied. Rather, Searle is given feedback on whether the outputs are correct. He then has to learn to make the right associations between inputs and outputs. Again, semantic meanings play no role.

According to Searle, what is present in a human thinker but absent in both the Chinese Room and a digital computer is *intrinsic intentionality*. A mental state exhibits intrinsic intentionality if it has semantic properties that are determined by the state alone, irrespective of its relations to anything else. Symbols in a computer may have semantic properties, but only extrinsically, as a matter of what they represent. But such extrinsic semantic properties are not 'available' to the system; they can make no difference to what it does. There is a long history of philosophers denying the existence of intrinsic intentionality, and a much longer history of philosophers insisting upon it.[2] (We'll return in the next section to the question of what could motivate an affirmative stance.)

In other work, Searle explicitly connects intrinsic intentionality to consciousness (Searle, 1992; see also Horgan, 2013). He points out that conscious intentional states present their referents 'under aspects'. That is, there is some way that you perceive a thing or think about it. And these aspects are comprehended by the mind, not merely symbolically represented. The idea is that consciousness is the exclusive means whereby subjects comprehend intentional contents. So, according to Searle, the problem of consciousness is related to the problem of semantic irrelevance in the following way: unless a system is conscious, it can't understand what its symbols mean, and if it can't understand what its symbols mean, it cannot perform any of its functions comprehendingly, in the way that human beings do.

Searle seems to assume that artificial systems are not conscious. Could he be wrong about this? Searle's own view is that consciousness reduces to biology rather than to functional structure, so he denies that functional structure as such is sufficient for consciousness. But this view is by no means universally held. Suppose, for example, that the global workspace theory correctly delimits the functional conditions sufficient for phenomenal consciousness. If we could build an artificial system that exemplifies the global workspace architecture, then perhaps such a system would comprehend meanings, and the problem of semantic irrelevance would be solved. Some experts think that this scenario is not too far off (see Dehaene, Lau and Kouider,

[2] For a denial, see Putnam (1981), and for an argument in its favour, see BonJour (1998). Denials go back at least as far as Wittgenstein. Something like a conception of intrinsic intentionality can be found in Aristotle and his medieval successors, and then again in modern thinkers such as Franz Brentano, Gottlob Frege, and Bertrand Russell.

2018). Or perhaps the global workspace theory exaggerates the functional complexity needed to underwrite phenomenal consciousness. If so, the age of conscious machines may have already dawned (see Carter *et al*., 2018).

But, even assuming that machines cannot be conscious, Searle's pessimistic case is incomplete. To say that humans exhibit a psychological feature that artificial systems lack is not yet to show that this feature is functionally necessary for genuine intelligence. First of all, it could turn out that intrinsic intentionality is epiphenomenal in humans. That is, even if conscious mental states have intrinsic intentionality, it may yet be those states' syntactic properties — say, the neural activation pattern with which they are correlated — that determines their functional role. This is, after all, a standard functionalist view of mental causation.

Second, even granting that semantic properties as such play a role in human intelligence, it does not follow they must play the same role in any intelligent system. Perhaps there are purely syntactic forms of cognitive processing that are just as powerful. After all, if a system can pass as a Chinese speaker without comprehending any meanings — a possibility granted by Searle's thought experiment — why not conclude that the causal relevance of semantic properties is less important to genuine intelligence than we thought?

5. Consciousness and Non-Syntactic Inferences

Each in his own way, both Baars and Searle draw a connection between consciousness, on the one hand, and the remediation of a specific functional deficit in artificial systems, on the other. (For Baars, the relevant deficit is the failure of the global availability of relevant information, and for Searle, the relevant deficit is the failure of cognitive relevance of semantic properties.) But neither completes his case. I now take up the slack. I will argue that the problems of global relevance and semantic irrelevance are serious, and that what makes them serious is the essential role that consciousness plays in genuine rationality. I claim that there are inferences, indispensable to genuine rationality, that cannot be reduced to the mechanical transformation of symbols, on pain of vitiating the epistemic status of those inferences. I call them 'non-syntactic inferences' for short.

I will discuss three types of non-syntactic inference: (1) inferences based on conceptual entailments and exclusions; (2) basic logical inferences; and (3) explanatory inferences.

5.1. Inferences based on conceptual entailments/exclusions

Among the most boringly obvious and cognitively fundamental inferences we make are inferences based on entailment- and exclusion-relations that hold among thought-contents. For example, if I know that x is red, I can infer that x is coloured. Similarly, if I know that x is red, then I can infer that x is not (completely) green. Note that I need no linking premise in order to make these inferences (e.g. 'if x is red then x is not completely green'). Consequently, these inferences are not licensed by any formal rule. What licenses them is my grasp of the properties of redness and greenness and of colour in general.

Grasp of such conceptual relationships is not limited to the case of colours. Rather, part of concept-possession is being able to locate contents on a conceptual 'map'. For example, knowing that x is a kite means knowing that x is a kind of toy, and thus not a type of planet or organism or social group; that it is distinct from a basketball; and so forth. (This dimension of concept-possession is known to philosophers as 'cognitive significance'.) One must have this ability for the world to exhibit even a minimal degree of intelligibility.

5.2. Basic logical inferences

In his classic article, 'What the Tortoise Said to Achilles', Lewis Carroll (1895) shows that no amount of propositional knowledge is psychologically sufficient for making rational inferences. Carroll imagines a thinker who assents to two propositions, A and B, and to the conditional *If A and B, then Z*, but who does not 'see' that Z follows. Simply adding the relevant inference rule as a premise will not help this person. She can grant that *If, If A and B, then Z, then Z*, but still wonder whether Z. She fails to grasp the rational connection that holds among the two propositions. The addition of more propositions to her belief-set will not remedy the situation.

But we are indeed able to grasp rational connections between propositions. Here is how Thomas Nagel characterizes the ability:

> Suppose I observe a contradiction among my beliefs and 'see' that I must give up at least one of them. (I am driving south in the early morning, and the sun rises on my right.) In that case, I see that the contradictory beliefs cannot all be true, and I see it simply because it is the case. I grasp it directly…
>
> In ordinary perception, we are like mechanisms governed by a (roughly) truth-preserving algorithm. But when we reason, we are like a mechanism that can see that the algorithm it follows is truth-preserving.

> Something has happened that has gotten our minds into immediate contact with the rational order of the world. (Nagel, 2012, p. 83)

There are two ways to characterize this ‘contact with the rational order of the world’. First, it might be that we can grasp rational connections among purely formal structures — e.g. we see that inference rules are valid — and then we grasp that particular sets of propositions are substitution-instances of those formal structures. Strikingly, this means that even syntactic inferences presuppose a type of understanding that extends beyond the syntactic. But Nagel’s compelling example of ‘just seeing’ that his beliefs are contradictory suggests that our inferential ability need not take this shape. More plausibly, our grasp of rational connections occur, in the first instance, in the ‘material mode’ rather than the ‘formal’. That is, we see that some *particular* proposition entails or excludes another *particular* proposition. Reasoning in the ‘formal mode’ is an abstraction out of reasoning in the ‘material mode’, which comes first epistemically and psychologically.

5.3. Explanatory inferences

Our ability to grasp rational connections among propositions extends beyond inferential connections. It also includes *explanatory* connections. Explanations come in many forms, so there are many species of the ability I have in mind. Crucial for my purposes is that understanding the explanatory relationship between two propositions is more than believing (or even knowing) some proposition that expresses that explanatory relationship. The key difference is that *understanding why* is more epistemically generative than *knowing why*. It confers on a subject a suite of abilities that Alison Hills calls ‘cognitive control’, which she explicates as follows:

> If you understand why p (and q is why p), then you believe that p and that q is why p and in the right sort of circumstances you can successfully:
> (i) follow some explanation of why p given by someone else.
> (ii) explain why p in your own words.
> (iii) draw the conclusion that p (or that probably p) from the information that q.
> (iv) draw the conclusion that p′ (or that probably p′) from the information that q′.
> (where p′ and q′ are similar to but not identical to p and q).
> (v) given the information that p, give the right explanation, q.

> (vi) given the information that p′, give the right explanation, q′. (Hills 2016, p. 663)

She illustrates cognitive control via the following case.

> Suppose that you know why giving money to charity is right, namely because we owe assistance to the very needy. You were told this by your parents. You understand what the statement means. You believe it and it is true. You have formed it in the right way to have knowledge (by testimony from sources you rightly believe to be reliable). But you won't necessarily yet have the abilities to make accurate judgements about other, similar cases — where you are aware that your sacrifice would be more significant, or that the needs you could meet are not pressing. And so on. So you can have knowledge why without cognitive control, and so without understanding why. (*ibid.*, pp. 669–70)

This is a case of understanding a moral explanation. Similar things could be said for understanding causal explanations, constitutive explanations, psychological explanations, teleological explanations, and so on. Hills' example involves grasping how one *general* moral claim explains another, but one can also grasp how some fact counts as a reason for performing some particular action or believing some particular proposition. Seeing something *as* a reason to act in a way, or to believe something, is to grasp an explanatory relationship between the reason and the candidate action or candidate belief. Inferences made on the basis of this explanatory relationship — from explanandum to explanans, or vice versa — are thus not purely syntactic.

5.4. Application to the three problems

So, we have identified three types of inference that involve grasping a connection between contents, a connection that syntax alone cannot capture. It is worth asking how the three types are related to each other. An intriguing possibility is that the first type grounds the second and third types. That is, if we grasp inferential and explanatory connections among propositions in the 'material mode' (in the first instance, at any rate), perhaps our grasp of relationships among propositions is parasitic on our grasp of relationships among *properties* — relationships of entailment, exclusion, and explanatory dependence (of various sorts).

In sum, to say that human cognizers engage in non-syntactic inference is to say the following: (a) we grasp the contents of our mental states;[3] (b) on the basis of our grasp, we *see rational connections* among those contents; (c) we make inferences that are guided by these rational connections.

In these observations, we find vindication for Searle's views about intrinsic intentionality. In the first place, our grasp of content would not be possible if the meanings of our mental representations were not directly present to our minds — which is to say that intrinsic intentionality is a real feature of our cognitive economy. In the second place, grasping is a form of conscious awareness. Thus, consciousness matters to rationality: conscious awareness is necessary to provide the rational insight required for our conceptual inferences and basic logical inferences to be warranted. Maybe (per my suggestion in the previous paragraph) a conscious grasp of sub-propositional meanings is sufficient to ground all three types of non-syntactic inference, or maybe there are additional semantic features of propositions that we grasp. Either way, it turns out that the problem of semantic irrelevance is a version of the problem of consciousness.

But the problem of global relevance also turns out to be a version of the problem of consciousness. Most of the literature on the problem has focused on the 'global' part of the problem, rather than the 'relevance' part of the problem. What is relevance? In the cases of interest to Dreyfus and Fodor, it is *explanatory* relevance. That is, abduction involves bringing to mind hypotheses that explain why something is or appears some way; practical deliberation involves bringing to mind considerations that explain why some course of action is recommendable or not. But human cognizers are consciously aware of such explanatory relations. Again: consciousness matters to rationality, because a conscious grasp of explanatory relations rationally guides abductive inferences.

The problems of global relevance and semantic irrelevance are pressing, it turns out, because they are special cases of the problem of consciousness. Consciousness licenses conceptual, logical, and explanatory inferences that do not make syntactic sense. To return to our earlier example: making the inference from 'x is red' to 'x is not

[3] Some of them, at any rate. Presumably our intentional reach exceeds our grasp. If we accept moderate externalism about content, then we can represent contents (natural kinds, concrete particulars, etc.) that we do not grasp.

completely green', a purely syntactic inference-generator would need an additional mediating premise:

x is red
↓
if x is red, then x is not completely green
↓
x is not completely green

For a conscious system, by contrast, it is as though the properties of redness and greenness themselves become part of the functional sequence, in lieu of this mediating premise:

x is red
↓
[grasp of red + grasp of green]
↓
x is not completely green

The way that consciousness licenses a non-syntactic inference is by making present to the mind the very thing that makes the inference rational (recall Nagel's point about our 'immediate contact with the rational order of the world'). And this means that consciousness plays the causal role it does in an irreducibly qualitative way, in virtue of its phenomenal character. No merely mechanical transition can replicate this role, while preserving the positive epistemic status of the inference.

6. Alternatives to Conscious Grasping: Reduction, Functional Work-Arounds, and Learning

I now discuss three objections to the argument just made. The first objection denies that consciousness plays the irreducible cognitive role I claim it does in human cognition. The second and third objections grant that it plays this role, but insist that the same functionality can be secured in other ways.

6.1. Objection 1: Grasping is reducible

I have been insisting that humans can make non-syntactic inferences on the basis of their grasp of rational connections among contents. But perhaps this is the wrong way to think about things. Perhaps grasping rational connections *just is* having the ability to make the inferences licensed by those connections. Such a move might be independently appealing, as it could it serve to reduce the mysterious phenomenon of grasping to something less mysterious.[4] But in the present context, it would also preserve a purely mechanical conception of rational inference.

The biggest problem with the reductive move is that it fails to preserve the epistemic status of the relevant inferences. My disposition to judge that x is not green from my knowledge that x is red is not epistemically basic. Rather, I am disposed to make this inference because the conceptual connection that I intuit licenses it. The reductive theory of grasping does not capture the sense in which non-syntactic inferences are rationally *guided*.

There is an obvious analogy with the phenomenon of blindsight. Just as there is an epistemic difference between making perceptual judgments on the basis of my perceptual experiences vs. making them, however reliably, in the absence of those experiences, so there is an epistemic difference between making non-syntactic inferences on the basis of my conscious grasp of rational connections among contents vs. making those same inferences, however reliably, in the absence of any such conscious grasp.[5] Grasping rational connections is thus explanatorily prior to the inferences rationalized by those connections, contra the reductive theory.

6.2. Objection 2: Grasping can be replaced with inferential rules or procedures

The second objection grants that consciousness plays the role in human cognition that I have been describing, but proposes functional work-arounds in purely syntactic systems. In other words, even if it would be debilitating to a *human* cognizer not to be able to consciously grasp connections among contents, differently constructed cognitive systems could function just as well.

There are two possible strategies here. The first is to add to the system explicit inference rules that reflect rational connections among

[4] In a similar vein, Bourget (2017) discusses two rival accounts of what it is to grasp propositional contents: the phenomenal theory and the inferential theory. The inferential theory of propositional grasping is analogous to the reductive strategy I have been discussing. Bourget suggests that the reductive move is almost irresistible to theorists who wish to give some sort of account or other of the phenomenon of grasping.

[5] Indeed, blindsight is probably in better epistemic shape, because a blindseer's judgments are at least causally dependent on their truthmakers. See Bengson (2015a,b) on the close analogy between perceptual experience and rational insight.

contents. In effect, non-syntactic inferences are replaced with syntactic ones. For example, where a human cognizer can just see that redness and greenness exclude each other, an artificial intelligence could be programmed with an inference rule that captures this exclusion-relationship. Or better, the artificial intelligence could be programmed to treat colour representations as points in a geometrical space, in which case rules of mathematics could serve the same purpose.

The second strategy is to build in *procedures* — epistemically basic causal dispositions — rather than explicit rules, that generate the same inferential results given the same premises.

I think it is fair to say that trends in AI favour the second option. Rule-governed computation is serial and resource-heavy, whereas procedures are 'quick and dirty' and occur in parallel. (Recall that Dreyfus's target is rule-governed computation as a viable model of human cognition.) Moreover, rule-governed computation must be implemented by procedures anyway. After all, one of the lessons from the Lewis Carroll problem is that adding explicit rules doesn't guarantee inference. As Hugo Mercier and Dan Sperber put the point: 'The representation of a regularity doesn't *do* anything all by itself, but it provides a premise that may be exploited by a variety of inferential procedures. A dedicated procedure *does something*: given an appropriate input, it produces an inferential output' (Mercier and Sperber, 2017, p. 87).[6]

Here the advocate of strong AI will point out just how much of *human* cognition functions automatically, via the executing of procedures. Recall Dreyfus's initial example:

> Mary saw a dog in the window. She wanted it.
> Mary saw a dog in the window. She smashed it.

Disambiguation of such sentences does not require that one consciously 'try out' various interpretations. That is, one does not first grasp an explanatory relationship between the proposition <Mary saw

[6] Hence the appeal of an architecture along the lines of the global workspace: in a centralized hub, serial, rule-based processing does occur, but most of the work is happening procedurally in the modules, operating in parallel. Now, Mercier and Sperber are advocates of the 'massive modularity thesis', according to which there is no central processer, just modules. This would be an endorsement of the second strategy 'all the way down'. Whether their view provides a genuine rival to a global workspace architecture has been questioned; see Chater and Oaksford (2018).

a dog> and <Mary wanted a dog>; then *fail* to grasp an explanatory relationship between <Mary saw a dog> and <Mary smashed a dog>; and finally, on the basis of the grasped explanatory relationship, make an inference to the correct disambiguation. Not only is the relevant inference, if made explicit, much more complicated than this (after all, one *can* smash a dog), but in normal circumstances the disambiguation occurs prior to conscious awareness of the stimulus. Once presented with an alternative disambiguation, one could perhaps say something about why the interpretation that 'naturally' occurred to one made sense. But the selection occurred without recourse to any such reasoning. Similar things could be said regarding practical deliberation. To a good bus driver, it will just *seem obvious* that she ought to take a passenger with a medical emergency to a hospital. She might be able to say something about why the situation warranted this course of action, but not because she consciously grasped an explanatory relation *before* deciding what to do.

And if abductive inference and practical deliberation can occur preconsciously, surely conceptual inferences and basic logical inferences could as well. To give just one example: developing one's ability to do 'mental math' is not a matter of improving one's grasp of mathematical relationships. It is, rather, a matter of developing the ability to make such inferences automatically. To put the point generally: expertise is usually a matter of being able to do things *without* consciously thinking about them, and expertise in making inferences is no exception. One might even suggest that the more a cognitive system depends on the execution of procedures and less on the implantation of rules, the more expert it is.

But we have already mentioned in passing one key difference that the ability to consciously grasp rational connections makes: a human can 'make sense' of her inferences after the fact, whereas a purely syntactic system cannot. It is this ability that confers on a human's inferences the positive epistemic status that they have. In many cases, one's conscious grasp of the relevant inferential connection is causally prior to the formation of non-conscious inferential abilities: first one grasps a connection, then one begins to exploit that grasp in the inferences one makes explicitly, and then, finally, such inferences become automatic. Non-conscious inferences then inherit their positive epistemic status from the conscious graspings that preceded them. But even where this is not the case — even if a non-conscious inferential ability is innate, for example — one's ability to 'check one's work', to appeal to the higher court of conscious rationality, makes one's conscious grasp *epistemically* prior to one's non-conscious inferences. Not so with machines whose inferential procedures are epistemically basic.

Thus, the problem with these functional work-arounds is the same as the problem with reduction, *viz*. they fail to preserve the positive epistemic status of the system's inferences. But, an engineer will ask, does it matter? If an artificial system's inferential dispositions *track* rationality, are not its inferences the epistemic equals of inferences that are guided by a grasp of the relevant inferential connections?

Surely not. For one thing, such a system is not able to correct its own errors. But even an artificial system that made no errors would be the epistemic inferior of a human, for the following reason: *that which makes its inferences rational is no part of the system itself.* For, consider a system that is reliably disposed to believe Z in circumstances of believing premises of a valid argument for Z. That is not the same thing as being rationally guided by the premises. The norms of rationality are (to reiterate) deontic, not consequentialist. Systems that are unable to grasp basic inferential connections may be perfectly reliable, but they are not guided by reason. They are fundamentally a-rational. They are, at best, epistemic appendages: powerful, automated extensions of the epistemic abilities of their conscious users.

6.3. Objection 3: Grasping can be replaced with rules or procedures that are autonomously learned

Artificial systems make inferences automatically, without being guided by a conscious grasp of the relevant rational connections. Humans also make inferences automatically, though often these automatic processes were formed on the basis of prior conscious grasping. This suggests that aetiology matters for the epistemic status of an inferential disposition. The aetiology of the inferential dispositions of an artificial system includes *human* rationality, either in the form of the explicit programming of rules or procedures (in the case of classical architecture) or in the context of the 'training' of a connectionist network — wherein the programmer tells the system what the output should be and lets the system construct its own procedure to get there. It is obvious, in these cases, that the artificial system is epistemically parasitic on its human programmer. But what if it could develop its inferential dispositions in a different way? What if it could learn to make reliable inferences without being taught to do so by a human? Perhaps such a system could be the functional *and epistemic* equal of a human intelligence.

We must imagine a system that initially makes inferences ‘willy-nilly’. The question at issue is this. Could such a system learn, ‘from scratch’, and not from an instructor but from the world itself, to reliably make conceptual inferences, basic logical inferences, and explanatory inferences? I will argue that the answer is no. For such a learning process presupposes the very thing that is to be learned.

Consider conceptual inferences first. We need to imagine a system whose representations lack ‘cognitive significance’. It has no idea which pairs of its representations denote distinct, identical, mutually excluding, or mutually entailing properties. Such a system could never learn to make conceptual distinctions. Suppose, for example, that such a system were tasked with determining whether two of its representations, ‘A’ and ‘B’, denoted the same content or different contents. To correctly infer that $A \neq B$, it would need to be in possession of two premises: first, a premise that states that A has a property that B lacks, and then a conditional expressing the indiscernability of identicals, as follows.

P1. $\exists F\ (FA \ \&\ {\sim}FB)$
P2. $\forall x \forall y\ (x=y \rightarrow \ \forall F(Fx \leftrightarrow Fy))$

(P1 can be read as ‘There is a property F such that A is F and B is not F’; or, more colloquially, ‘A has a property that B lacks.’ P2 can be read as ‘For all x and all y, if x and y are identical, then for any property F, x is F if and only if y is F’; or, more colloquially, ‘If x and y are identical, then any property x has y also has, and vice versa.’) But how could the system come to know P1? Suppose it learns that A is F. To know P1, it would need to know that B is not F. Suppose it has an exhaustive list of B’s properties, G_1–G_n. Unless it is in a position to know that F is not identical to any of G_1–G_n, it is not in a position to acquire P1 on the basis of this exhaustive survey. So, in order to learn that $A \neq B$, it would first need to learn that $F \neq G_1$, $F \neq G_2$, $F \neq G_3$, etc. And this would require a prior inference involving the indiscernibility of identicals, and each of these inferences would rely on a premise akin to P1. Thus, on pain of infinite regress, cognitive significance must be built into the system somehow. Similar things can be said with respect to learning empirically that ‘two’ contents are in fact identical.[7] Of course, with a little conceptual knowledge, the system might be able to learn a lot of it. But that much is true of human cognizers, as well. The point is that the system cannot start from scratch; some of its conceptual knowledge must be given — by its human programmer.

Nor could a system acquire the ability to make basic logical inferences or abductive inferences without a modicum of these abilities pre-programmed. Consider basic logical inferences. There is, of course, a way to determine which sentences in a formal system are theorems of that system, *viz*. the method of *reductio ad absurdum*. But — and this is an old point — it is impossible to carry out such a procedure without first knowing some inference rules.

Or again, consider abductive inference. Could a system learn which of its representata are explanatorily relevant to one another? The procedure would involve the formation of meta-hypotheses — the hypothesis that hypothesis H is relevant to explanandum E. The system would then need to test this hypothesis. (To make this suggestion more vivid: certain contextual clues are relevant to the meaning of utterances, and others are not. To determine which are and which are not, the system forms the hypothesis that contextual clues of a certain type — say, the utterer's tone of voice — are relevant. The system then tries bringing tone of voice to bear on the disambiguation of utterances, and then evaluates whether the strategy seems to be working.) But the problem is obvious. Testing the hypothesis requires that the system be able to tell which results of the trial-and-error process are relevant and which are not. Only a system that already understands basic explanatory relationships can test hypotheses about explanatorily relationships.

In sum: human cognizers make inferences that are guided by the conscious grasping of non-syntactic relationships among contents. And there is no epistemically equivalent substitute. Pre-programmed rules or procedures that track rationality inherit their epistemic *bona fides* from the programmer's rational capacities. But if rules or procedures are not pre-programmed, the system has no way to learn them.

[7] See Woodward (2018) for a more extended discussion of this regress principle.

7. Conclusion

The last two sections have taken us deep into abstruse epistemological difficulties. Let's come up for air.

My topic has been three sources of pessimism about the prospect of literal artificial intelligence: the problem of consciousness (per Lister), the problem of global relevance (per Dreyfus and Fodor), and the problem of semantic irrelevance (per Searle). The problem of semantic irrelevance turns out to be a special case of the problem of consciousness, because it is consciousness alone that renders semantic content intrinsic to mental states. Only if contents are intrinsic can a subject grasp rational connections among them. Grasping such connections licenses non-syntactic inferences that form the beating heart of rationality.

The problem of global relevance also turns out to be a special case of the problem of consciousness. The problem of global relevance is partly solvable simply by mimicking the functional structure of human consciousness, a structure known as the 'global workspace' architecture, which consists in a hybrid of serial, domain-general processing and parallel, modular processing. Baars is right that consciousness is needed for global psychological processes, but not simply because consciousness happens to play the role of integrating information from specialized processers. Rather, only consciousness can supply the right sort of insight into explanatory relations that is necessary for the central processor to be rationally guided in its abductions and practical deliberations.

How, then, shall we adjudicate the dispute between optimists and pessimists? We have certainly not shown that artificial intelligence is impossible. If such a conclusion were the argumentative ambitions of the likes of Dreyfus, Fodor, or Searle, then their ambitions remain unfulfilled. Moreover, none of the foregoing has suggested that there are cognitive feats that only conscious systems can perform. Purely syntactic systems, we have shown, cannot be the epistemic equals of conscious systems. But, for all we've said, their performance on any given cognitive task might be just as good or better.

At the same time, we have identified rather narrow constraints on any genuinely intelligent artificial system. First, such a system would have to be conscious. Second, its conscious states would have to be of a certain sort, *viz.* they would need to afford the system a direct grasp of the contents of some of its representational states. Third, these conscious states would have to play a peculiar functional role, *viz.* they would need to mediate between inferential inputs and outputs *in virtue of their qualitative character*. Then, and only then, could the system engage in rational, non-syntactic inference (which, as the Lewis Carroll problem makes clear, is an ingredient in rational inference of every sort). I defer to AI researchers to determine just how optimistic we should be about the satisfiability of these constraints.

But there is a more general lesson. Contemporary psychologists tend to conceive of the cognitive role of consciousness in terms of those psychological functions that must be performed consciously if performed at all. Many candidate functions have been proposed (I borrow here Tom Polger's list): flexible behaviour, creativity, communication, mental rehearsal, self-knowledge, mentalistic language, and self-awareness of a special sort (Polger, 2017, p. 82). If we are understanding 'psychological function' in a purely causal way, then it is at best an anthropological law that consciousness does these things, not a psychological law, let alone a metaphysical law. As Polger puts the point: 'Consciousness may, of course, be necessary for our way of doing things. But that will not show that consciousness had to occur unless it is also necessary that we evolved to be as we are — which surely it is not' (*ibid.*, p. 90). Researchers who conceive of the cognitive role of consciousness in purely causal terms will never capture their quarry: they look for psychological necessities on terrain where only engineering contingencies can be found.

But this just shows that we should not understand the cognitive role of consciousness in causal terms, but also in epistemic terms. Consciousness is essential to cognition not in what it *causes* but what it *licenses*, *viz*. non-syntactic inference. What psychological subjects gain by being conscious is an awareness of the rational connections among things. It is no specific type of rational process that consciousness underwrites, but rather the very possibility of any process's counting as rational at all.

Acknowledgments

This project was made possible through the support of the Henry Center for Theological Understanding, funded by a grant from the John Templeton Foundation. The opinions expressed in this publication are those of the author and do not necessarily reflect the views of the John Templeton Foundation.

References

Baars, B.J. (1988) *A Cognitive Theory of Consciousness*, Cambridge: Cambridge University Press.

Bengson, J. (2015a) Grasping the third realm, *Oxford Studies in Epistemology*, **5**, pp. 1–38.

Bengson, J. (2015b) The intellectual given, *Mind*, **124** (495), pp. 707–760.

BonJour, L. (1998) *In Defense of Pure Reason: A Rationalist Account of A Priori Justification*, Cambridge: Cambridge University Press.

Bourget, D. (2017) The role of consciousness in grasping and understanding, *Philosophy and Phenomenological Research*, **95** (2), pp. 285–318.

Carroll, L. (1895) What the tortoise said to Achilles, *Mind*, **4** (14), pp. 278–280.

Carter, O., Hohwy, J., Van Boxtel, J., Lamme, V., Block, N., Koch, C. & Tsuchiya, N. (2018) Conscious machines: Defining questions, *Science*, **359** (6374), p. 400. doi: 10.1126/science.aar4163

Chater, N. & Oaksford, M. (2018) The enigma is not entirely dispelled: A review of Mercier and Sperber's *The Enigma of Reason*, *Mind & Language*, **33** (5), pp. 525–532.

Churchland, P. (2013) *Touching a Nerve: Our Brains, Our Selves*, New York: W.W. Norton & Company.

Crane, T. (2015) *The Mechanical Mind*, New York: Routledge.

Deco, G., Vidaurre, D. & Kringelbach, M.L. (2021) Revisiting the global workspace orchestrating the hierarchical organization of the human brain, *Nature Human Behaviour*, **5** (4), pp. 497–511.

Dehaene, S. & Naccache, L. (2001) Towards a cognitive neuroscience of consciousness: Basic evidence and a workspace framework, *Cognition*, **79** (1–2), pp. 1–37.

Dehaene, S., Lau, H. & Kouider, S. (2018) Response, *Science*, **359** (6374), pp. 400–402.

Dreyfus, H.L. (1965) *Alchemy and Artificial Intelligence*, No. P-3244, Santa Monica, CA: Rand Corp.

Dreyfus, H.L. (1979) *What Computers Can't Do: The Limits of Artificial Intelligence*, vol. 1972, New York: Harper & Row.

Fodor, J.A. (1983) *The Modularity of Mind*, Cambridge, MA: MIT Press.

Fodor, J.A. (2000) *The Mind Doesn't Work That Way: The Scope and Limits of Computational Psychology*, Cambridge, MA: MIT Press.

Hills, A. (2016) Understanding why, *Nous*, **49** (2), pp. 661–688.

Horgan, T. (2013) Original intentionality is phenomenal intentionality, *The Monist*, **96** (2), pp. 232–251.

Mercier, H. & Sperber, D. (2017) *The Enigma of Reason*, Cambridge, MA: Harvard University Press.

Nagel, T. (2012) *Mind and Cosmos: Why the Materialist Neo-Darwinian Conception of Nature is Almost Certainly False*, New York: Oxford University Press.

Polger, T. (2017) Rethinking the evolution of consciousness, in Velmans, M. & Schneider, S. (eds.) *The Blackwell Companion to Consciousness*, pp. 180–199, Oxford: Blackwell.

Putnam, H. (1981) *Reason, Truth and History*, vol. 3, Cambridge: Cambridge University Press.

Searle, J.R. (1980) Minds, brains, and programs, *Behavioral and Brain Sciences*, **3** (3), pp. 417–424.

Searle, J.R. (1992) *The Rediscovery of the Mind*, Cambridge, MA: MIT Press.

Shanahan, M. & Baars, B. (2005) Applying global workspace theory to the frame problem, *Cognition*, **98** (2), pp. 157–176.

Turing, A. (1950/2010) Computing machinery and intelligence, reprinted in Morton, P.A. (ed.) *A Historical Introduction to the Philosophy of Mind*, 2nd ed., pp. 330–346, Peterborough, ON: Broadview Press.

Woodward, P. (2018) A posteriori physicalism and the discrimination of properties, *Acta Analytica*, **33** (1), pp. 121–143.

Paper received October 2021; revised February 2022.