Journal of Experimental and Theoretical Artificial Intelligence 12 (3):263-277 (2000)

To appear in the *Journal of Experimental & Theoretical Artificial Intelligence* (JETAI)

Consciousness, Intentionality, and Intelligence: Some Foundational Issues for Artificial Intelligence

MURAT AYDEDE
The University of Chicago
Department of Philosophy
1010 East 59th Street
Chicago, IL 60637
(773) 702-8513 (office)
(773) 702-9861 (fax)
m-aydede@uchicago.edu

GÜVEN GÜZELDERE
Duke University
Department of Philosophy
201 West Duke Building, Box 90743
Durham, NC 27708
(919) 660-3068 (office)
(919) 660-3060 (fax)
guven.guzeldere@duke.edu

ABSTRACT: We present three fundamental questions concerning minds. These are about consciousness, intentionality and intelligence. After we present the fundamental framework that has shaped both the philosophy of mind and the Artificial Intelligence research in the last forty years or so regarding the last two questions, we turn to consciousness, whose study still seems evasive to both communities. After briefly illustrating why and how phenomenal consciousness is puzzling, we propose a theoretical diagnosis of the problem and present a framework within which further research would yield a solution. Our diagnosis is that the puzzle stems from a peculiar dual epistemic access to phenomenal aspects (qualia) of our conscious experiences. We present an account of concept formation such that both the phenomenal concepts (like the concepts, RED and SWEET) and the introspective concepts (like the concepts, EXPERIENCING RED and TASTING SWEET) are acquired from a first-person perspective as opposed to the third-person one (the standard concept formation strategy about objective features). We explain the first-person perspective in information-theoretic and computational terms.

> Nature (the Art whereby God hath made and governes the World) is by the Art of man, as in many other things, so in this also imitated, that it can make an Artificial Animal. For seeing life is but a motion of Limbs, the beginning whereof is in some principall part within; why may we not say, that all Automata (Engines that move themselves by springs and wheels as doth a watch) have an artificiall life? For what is the Heart, but a Spring; and the Nerves but so many Strings; and the Joynts, but so many Wheeles, giving motion to the whole Body, such as was intended by the Artificer? Art goes yet further, imitating that Rationall and most excellent worke of Nature, Man. (Hobbes 1651: 81)

So declared Thomas Hobbes in 1651 in the Introduction to his well-known work, Leviathan, published one year after René Descartes' death. Descartes was also interested in mechanical explanations of bodily processes and organic life. In fact,

on the basis of his neuroanatomical and physiological studies, as well as philosophical arguments, Descartes had already argued that human and animal bodies could be mechanically understood as complicated and intricately designed machines (Descartes 1664). What differentiated Descartes from Hobbes lay in Descartes' belief that human beings, unlike non-human animals, were not merely bodies; they were unions of material bodies and immaterial souls. The immaterial soul was necessary for Descartes to explain the peculiar capacities and activities of the human mind. As such, materialist mechanical explanations could never be sufficient to account for the whole human being.[1]

The fundamental assumption of Artificial Intelligence (AI) as a research program is that human minds operate on computational principles, and its grand goal is to build material artifacts that genuinely possess the very same mental capacities that human beings have. As John Haugeland puts it, 'we are really interested in AI as part of the theory that people are computers' (Haugeland 1985: 5–6). If so, in order for the project of AI to have any hopes of accomplishing its grand goal, it has to rely on an entirely materialist framework. The important and relevant theoretical question, which connects foundational considerations of Philosophy with the empirical considerations of AI research, is, then, whether and how a materialist account of the mind can be given. This is the question we will explore in this essay, in light of the most recent developments in contemporary philosophy of mind.

1 Conceptual Foundations

One of the central tenets of contemporary philosophy of mind, which fits well with the general framework of AI research, lies in its commitment to an ongoing research program in "naturalizing the mind". The naturalization program in philosophy of mind is an attempt to provide a theoretical framework in which the mind can naturally be seen as part of the physical world without postulating irreducibly psychic entities, events, processes, or properties for its explanation.

Jerry Fodor, one of the most influential figures in present-day philosophy of mind, once identified the following three questions as the major open problems in the field:

> How could anything material have conscious states? How could anything material have semantical properties? How could anything material be rational? (where this means something like: how could the state transitions of a physical system preserve semantical properties?). (Fodor 1991: 285, Reply to Devitt)

Fodor's own theory, the computational/representational theory of mind (CRTM), is a full-blown attempt to give a naturalistic answer to the third question, and an attempt to solve at least part of the problem underlying the second one. But it is almost silent about the first.[2] This discrepancy is not peculiar to Fodor's work, however. Many contemporary philosophers believe that while CRTM can in principle give a full account of thinking, believing, planning, intending, judging, and the like, the explanation of qualitative aspects of the mind — such as colour

sensations, feelings of cold and warmth, and tickles and pains (and perhaps feelings of sadness, anger, and joy, as well) — lies beyond the reaches of any such theory.

Whether or not AI researchers agree with philosophers on the discrepancy between the prospects of explaining consciousness versus explaining intentionality and rationality, it is a fact that most of the work in AI research so far has heavily focused on the latter issues, and hardly ever on the former. Herbert Simon draws this distinction vividly in the preface to the discussion of his thesis that 'a man, viewed as a behaving system, is quite simple' and that 'intelligence (as computation) is the work of symbol systems' (Simon 1969: 65, 28). Simon declares:

> Instead of trying to consider the “whole man”, fully equipped with glands and viscera, I should like to limit the discussion to Homo sapiens, "thinking man". I myself believe that the hypothesis holds even for the whole man, but it may be more prudent to divide the difficulties at the outset, and analyze only cognition rather than behavior in general. (Simon 1969: 65)

Our goal in this paper is not to show how CRTM succeeds or fails in answering Fodor's three questions. Rather, we would like to highlight the fact that there is in fact a discrepancy between the first question, on the one hand, and the second and third questions on the other, and then to point out that this discrepancy appears for both good and bad reasons. The good reasons have to do with a crucial difference between purely intentional cognitive states, such as beliefs and desires, and phenomenally conscious states, such as sensations. The bad reasons have to do with an implicit assumption that theorizing about intentional cognitive states can never illuminate questions about phenomenally conscious states.

The structure of the paper is as follows. First we explain the difference between merely cognitive mental states and conscious mental states. Then we sketch how CRTM construes propositional attitudes and promises to answer Fodor's second and third questions. Finally, we focus our discussion on the problem of conscious states, and eventually propose a new approach to this problem, which draws on the resources of CRTM.

2 The Computational/Representational Theory of Mind (CRTM)

It is common practice in everyday life to attribute a variety of mental states — beliefs, desires, hopes, fears, regrets, expectations, etc. — to people (and sometimes to non-human animals and even certain artifacts) to make sense of their behaviour. Philosophers standardly call such states propositional attitudes, because they seem to be mental attitudes towards propositions. For example, in the case of Pat's believing that Istanbul is a beautiful city, Pat's belief is construed as a relation between an agent (Pat) and a proposition, conceived as some sort of abstract object and expressed by the complement sentence 'Istanbul is a beautiful city'.

It is also common practice to regard beliefs as standing in various semantic, evidential, and inferential relations to one another. Thus, we expect that, if John believes that all police officers are corrupt and comes to believe that Smith's brother is a police officer, then other things being equal, John will come to believe that Smith's brother is corrupt. Notice that in this kind of discourse, it is <u>beliefs</u> that are claimed to stand in implication relations, not just <u>objects</u> of belief. Compare:

- What John believes is contradicted by what Smith believes, and confirmed by Alice's experience with police officers.

The situation here seems to be quite general with respect to other attitudes and other semantic and epistemic relations, such as logical equivalence, synonymy, and disconfirmation.

Finally, we can think of inference as a causal process: it is in virtue of his prior two beliefs that John now comes to have the third belief. For instance, John's first two beliefs cause the belief that Smith's brother is corrupt, but not the belief that Mary is corrupt, or the belief that two plus two is four. This is no accident, on the present view, for beliefs causally interact in ways sensitive to their content.

Most importantly, practical reasoning and the production of behaviour are typically responsive to the content of the beliefs and desires involved. If Alice believes that permitting the free use of marijuana will be beneficial, and she hopes that it will be so permitted one day, then whenever there is a public referendum as to whether marijuana use should be legalized and Alice believes that her vote can make a difference, other things being equal, she will typically form a desire to vote 'yes' in the referendum, and will vote accordingly.

Such means-ends reasoning is paradigmatically responsive to what is wanted and what is believed. It is because I believe that drunk driving is potentially life-threatening, and desire not to take a risk, that I form the desire to avoid alcohol at the party, which in turn is causally involved in my ensuing behaviour (driving sober). I formed the desire I did at that point — rather than, say, the desire to eat

chocolate ice cream — because of my prior attitudes: their content was relevant to the causal explanation of why I formed the particular desire to avoid alcohol and why I behaved the way I did. On this view, what is believed and desired appear to have overlapping parts — shared conceptual elements — and these are what the causal story underlying inference, practical reasoning, and the production of behaviour appeals to.

Thinking, practical reasoning and rational behaviour, therefore, all involve causally proceeding from states to states (and ultimately to behaviour) that would make semantic sense: the transitions among states must preserve some of their semantic properties to count as thinking. In the ideal case, this property would be the truth value of the states. But in most cases, any interesting intentional property like warrantedness, degree of confirmation, semantic coherence given a certain practical context like satisfaction of goals in a specific context, etc. would do. In general, it is hard to spell out what this requirement of "making sense" comes to. The intuitive idea, however, should be clear. Thinking is not proceeding from thoughts to thoughts in arbitrary fashion: thoughts that are causally connected are in some fashion semantically connected too. If this were not so, there would be little point and gain in thinking. This general phenomenon, the semantic coherence of causally connected thought processes, is what Fodor's third question is all about. CRTM is offered as a solution to this puzzle: how is thinking (and rational behaviour), conceived this way, physically possible?

In light of this brief exposition, let us now outline the Computational/Representational Theory of Mind (cf. Field 1978: 37, Fodor 1987: 17):

(A) **Representationalism:**

(1) Representational Theory of Thought:

For each propositional attitude A, there is a unique and distinct (i.e. dedicated)[3] psychological relation R and for all propositions P and subjects S, S As that P if and only if there is a mental representation #P# such that

(a) S bears R to #P#, and

(b) #P# means that P.

(2) Representational Theory of Thinking:

Mental processes, thinking in particular, consist of causal sequences of tokenings of mental representations.

(B) **Computationalism**: Mental representations, which, as per (A1), constitute the direct "objects" of propositional attitudes, belong to a representational or symbolic system which is such that (cf. Fodor and Pylyshyn 1988: 12–13)

(1) representations of the system have a combinatorial syntax and semantics: structurally complex (molecular) representations are systematically built up out of structurally simple (atomic) constituents, and the semantic content of a molecular representation is a function of the semantic content of its atomic constituents together with its syntactic/formal structure, and

(2) the operations on representations (constituting, as per (A2), the domain of mental processes) are causally sensitive to the syntactic/formal structure of representations defined by this combinatorial syntax.

(C) **Physicalist Functionalism**: Mental representations so characterized are functionally characterizable entities which are realized by physical properties of the subject of the attitudes (if the subject is an organism, then the realizing properties are presumably the neurophysiological properties in the brain or the central nervous system).

The relation R in (A), when (A) is combined with (B), should be understood as a computational/functional relation. The idea is that each attitude is identified with a characteristic computational/functional role played by the mental sentence that is the direct object of that kind of attitude. For instance, what makes a certain mental sentence an (occurrent) belief might be that it is characteristically the output of perceptual output systems and input to an inferential system that interacts decision-theoretically with desires to produce further sentences or actions. Or equivalently, we may think of belief sentences as those that are accessible only to certain sorts of computational operations. Similarly, desire sentences (and sentences for other attitudes) may be characterized by a different set of operations that jointly constitute a characteristic computational role for them. In the literature it is customary to use the metaphor of a "belief-box" (cf. Schiffer 1981) as a blanket term for whatever computational role belief sentences have in the mental economy of their hosts. (Similarly for "desire-box", etc.)

The two most important achievements of 20th century that are at the foundations of CRTM as well as most of modern Artificial Intelligence (AI) research and the so-called information processing approaches to cognition (practically almost all of contemporary cognitive psychology) are (i) the developments in modern

symbolic (formal) logic, and (ii) Alan Turing's idea of a Turing Machine and Turing computability. It is putting these two ideas together that gives CRTM its enormous explanatory power within a naturalistic framework. Modern logic showed that most of deductive reasoning can be formalized, i.e. most semantic relations among symbols can be entirely captured by the symbols' formal/syntactic properties and the relations among them. And, Turing showed, roughly, that if a process has a formally specifiable character then it can be mechanized. So we can appreciate the implications of (i) and (ii) for the philosophy of psychology in this way: if thinking consists in processing representations physically realized in the brain (in the way the internal data structures are realized in a computer) and these representations form a formal system, i.e. a language with its proper combinatorial syntax (and semantics) and a set of derivations rules formally defined over the syntactic features of those representations (allowing for specific but extremely powerful programs to be written in terms of them), then the problem of thinking (and rational action), as we described it above, can in principle be solved in completely naturalistic terms, thus the mystery surrounding how a physical device can ever have semantically coherent state transitions (processes) can be removed. Thus, given the commitment to naturalism, the hypothesis that the brain is a kind of computer trafficking in representations in virtue of their syntactic properties is the basic idea of CRTM and the AI vision of cognition.

Computers are environments in which symbols are manipulated in virtue of their formal features, but what is thus preserved are their semantic properties, hence the semantic coherence of symbolic processes. This is in virtue

of the mimicry or mirroring relation between the semantic and formal properties of symbols. As Dennett once put it in describing LOTH, we can view the thinking brain as a syntactically driven engine preserving semantic properties of its processes, i.e. driving a semantic engine.

To sum up: CRTM, as sketched above, provides a way of understanding how phenomena such as thoughts and beliefs, as well as thinking, decision making, practical reasoning and rational action, can be understood in a materialist framework that not only can explain human mentality in terms of bodily processes but also points to how they might be implemented in other physical systems, including artifactual ones (e.g. robots). This is how CRTM is theoretically equipped to tackle Fodor's second and third questions.[4] It does remain silent, however, when it comes to the first question, the question of consciousness. How is it that a physical system can come to have qualitative states — experience flashes of colours, feel pangs of jealousy, or enjoy the warmth of the afternoon sun? This is the problem to which we now turn.

3 The Problem of Phenomenal Consciousness: Experience

The problem of experience concerns the ontological status of the qualitative character of our experiences — their qualitative feel, or 'qualia' — of which we seem to be directly aware in introspection.[5] It is characterized here as a problem because, on the face of it, it is not clear how qualia could be entirely physical (e.g. some sort of entirely physical phenomena in the brain). But the puzzling character of qualia is a more general problem about our understanding of them, because if it is puzzling to

think of qualia in physical terms, it is no less puzzling to think of them in non-physical terms. The mystery remains even if physicalism is rejected. This aspect of the problem is brought out nicely by Jackson's so-called "Knowledge Argument".[6] Here is the thought-experimental set-up for the argument:

> Mary is confined to a black-and-white room, is educated through black-and-white books, and through lectures relayed on black-and-white television. In this way she learns everything there is to know about the physical nature of the world. She knows all the physical facts about us and our environment, in a wide sense of 'physical' which includes everything in completed physics, chemistry, and neurophysiology, and all there is to know about the causal and relational facts consequent upon all this, including of course functional roles. (Jackson 1986: 567)

Mary is released and sees for the first time a ripe tomato in good light, and comes to know what it is like to see red, something she allegedly did not know before, despite her omniscience with respect to physical facts. Jackson runs his argument thus (1986: 568):

> (1)' Mary (before her release) knows everything physical there is to know about other people.
>
> (2)' Mary (before her release) does not know everything there is to know about other people (because she learns something about them on her release).
>
> Therefore,

(3)' There are truths about other people (and herself) which escape the physical story.

According to Jackson, physicalism is the doctrine that the world consists entirely of physical facts. if that doctrine is correct, someone who knows all the physical facts knows all there is to know. According to Jackson, Mary comes to know a new fact upon seeing red for the first time, a fact which she did not know before; but since, by hypothesis, she already knew all the physical facts, the fact she comes to know cannot be physical. Hence, there are non-physical facts, and physicalism is false.

Jackson seems to think that in experience we encounter, are acquainted with, (instantiations of) non-physical properties. But if qualia are non-physical, it is hard to see how they could participate in the causal working of the physical world, which includes our bodies. According to (early) Jackson (1982), and many other anti-physicalists, qualia are epiphenomenal: they are caused (by physical events) but they don't cause anything, they are altogether causally inefficacious. So it seems that there is a heavy price to pay if physicalism is false. For the falsity of physicalism makes the mystery bigger, not smaller.

4 How to Approach the Problem

Although to some extent we share the sense of awe and mystery surrounding the philosophical problem of phenomenal consciousness described above, we are no mysterians about consciousness. In fact, we are optimistic about the prospects for a naturalistic solution. In what follows, we will indicate the grounds for this optimism, and describe in broad outline the theoretical tenets of a naturalistic

research program within which consciousness, and not just intentional cognitive states, can be explained. If we are right about how to pursue this research, the ultimate solution will be an interdisciplinary one, involving not only the relevant branches of neuroscience and psychology but also AI in a crucial way.

As we characterized the problem of consciousness above, a particular form of state consciousness becomes the focus of mystery. It is important to note at this juncture that there are two kinds of mental states that can be conscious: phenomenal states (sensory and emotional experiences , like pains, itches, seeing red, smelling coffee, and feeling depressed), and cognitive states with conceptual content (propositional attitudes like thoughts, beliefs, and desires). Although it is problematic how any such states could be conscious, the degree of mystery that attaches to both kinds is not the same. There is a sense that explaining what makes a thought conscious is easier than explaining what makes an experience conscious. Indeed, the sense of philosophical mystery always accompanies the latter and almost never the former.

For instance, McCarthy (1999) argues that making robots conscious is in principle within our grasp. However, it turns out that what McCarthy has in mind is robots' capacity to have conscious thoughts (propositional attitudes), not experiences. He seems to join the group of people who declare conscious experience a mystery, and as a result, he questions not only the possibility of robots' having conscious experiences, but also the desirability of producing robots with this capacity, assuming it were possible to do so. He seems to think that having conscious experiences is an option that robots with fully conscious thoughts could do without.

McCarthy, of course, is not alone in this regard. Some philosophers also think much along the same lines.[7]

What makes the problem of having conscious thoughts easier, or at least, seem easier? There is some consensus about what the shape of the right explanation would look like here. The answer offered relies on the existence of a particular kind of higher-order cognitive access to thoughts and other first-order cognitive states, which become conscious states in virtue of this access. On this view — what we will call the Higher-Order Representation (HOR) view — a thought (propositional attitude) is conscious if and only if it is the direct object of a representational state of the same mind.

There are two versions of HOR. One version, the Higher-Order Perception (HOP), view, takes the higher order access to be perception-like. The other, the Higher-Order Thought (HOT) view, takes this access to be more conceptually sophisticated, involving a higher-order thought about the mental state that is said to become conscious thereby. Although the philosophical tradition behind the first version, which regards consciousness as a kind of “inner sense”, is a long and venerable one, its present status is controversial. For there are grounds for doubting the existence of a kind of sense organ (having a status like that of the exteroceptive senses) dedicated to detecting mental events in one's mind — a "mind's eye", if you will. But even if HOP turns out to be right, it seems plausible that when applied to first-order thoughts, its ultimate utility would lie in its being a kind of intermediary to knowledge about first-order states. To the extent that knowledge is conceptually

articulated, the utility of HOP would consist in yielding HOTs about first-order states. We will return to HOP below and review its status again.[8]

Whether or not the HOT theory is correct as a view of state consciousness in general, it is clear that we can form thoughts about our own experiences and thoughts which have the relevant kind of directness and immediacy. Whatever the actual mechanisms of such a capacity are, a general outline of their account seems not very problematic given a certain picture of the cognitive mind, namely that of CRTM. In other words, once we have the general outline of a theory, like CRTM, of what it is to think thoughts, it seems trifling, at least from an engineering point of view, to add mechanisms for thinking about those thoughts. Indeed, McCarthy's (1999) suggestions seem to point in this very direction.

However, HOR theories seem to be less plausible when applied to experiences, for two reasons. First, it is less clear what the mechanisms underlying HORs about experiences are; and second, the mystery surrounding phenomenal consciousness seems to persist even when HOR accounts are in place. On the face of it, if it is puzzling to have experiences such that there is always something it is like to be in them for the experiencer, it seems equally puzzling to be told that what it is like to be in them is nothing over and above having HOR states about those experiences. How is this supposed to advance our understanding of the what-it-is-like aspect of experiences? We think that many in fact share this intuition that HOR accounts bypass the problem of phenomenal consciousness altogether without offering any insight or advance in our understanding.

Although we share this intuition with respect to many HOR accounts in the literature, we nevertheless think that there is a sense in which such accounts are on the right track. Regarding their plausibility as applied to experiences, the difficulty here stems from a comparative lack of understanding of the interface between sensory experiences and thoughts. If we take sensory modalities and our experiences in them as information channels opening windows to our environment en route to more central cognitive processing and behaviour appropriate to that environment, then it is plausible to take experiences as representational states that encode information about the environment (external as well as bodily) in a format different than the format in which thoughts encode information about that environment. Sometimes this difference is captured by saying that the representational content of sensations / experiences is encoded in analog (=non-conceptual) form whereas that of thoughts in digital (=conceptual) form. We have seen the outlines of what the digital / conceptual format comes down to in the case of thoughts within the framework of CRTM. Perhaps it is the lack of knowledge of what the analog format of sensory information comes to and the interface between these two sorts of format that creates the first problem we have just mentioned. But once the problem is put this way, it is clear that solution to it would come from doing more empirical (as well as foundational) research. In this regard, it seems likely that robotics research and recent work on embedded (situated) computation can in principle shed substantial light on the ultimate solution. We do not think this aspect of HOR views should be very problematic or mysterious.

The second way in which HOR accounts seem problematic when applied to experiences is more serious. Nonetheless, we also believe that once we have a correct diagnosis of what creates the mystery and a more sophisticated HOR account is given in light of this, it will become clear how the notion of HOR can add to our understanding of phenomenal consciousness. To state our diagnosis, we will now revisit the thought experiment involved in Jackson's anti-physicalist argument.

5 The Knowledge Argument Revisited

The diagnosis we would like to present is at the core of a certain group of materialist responses to Jackson's argument.[9] It consists in acknowledging that Mary does indeed come to learn something new and factual in character. however, what she learns is not a new fact, but rather a conceptually (epistemically) new way of relating to an "old" fact which she already knew under its objective physical description. Mary already knew what experiencing red is under its scientific description: she knew how red objects strike the retina, how the brain processes the retina's output in different areas of the visual cortex, and so on. Let us say, then, that Mary knew that

(1) experiencing red is sde_R,

where 'sde_R' stands for the complete scientific description of experiencing red. In this sense, Mary already had the necessary concept(s) expressed by 'sde_R'. Upon looking at a ripe tomato for the first time after her release, she comes to occupy a

certain experiential/brain state for the first time, to which she knows the description 'sde_R' applies. But now, consequent upon experiencing red for the first time, she also comes to acquire a new concept. She can now represent her experience thus:

(2) experiencing red is like this, (or, this is experiencing red)

where 'this' expresses the mental tokening of a certain perspectival concept she has just acquired that in turn expresses the same property expressed by 'sde_R'. It is important to be clear about what is new in Mary when she first experiences red. First, there is the objective property of redness physical objects possess (a certain class of surface spectral reflectances). Second, there is the visual experience of red, exp_R, which we will treat as a non-conceptual representation of objective redness. Finally, there is the experiential concept EXP_R Mary acquires consequent upon experiencing red. Schematically, the dependency relations look like this:[10]

redness <— exp_R <— EXP_R,

Part of what makes EXP_R perspectival is this: (i) necessarily, Mary could not have acquired EXP_R had she not had exp_R , and (ii) necessarily, EXP_R acquires its extension partly in virtue of standing in a special informational (direct causal/nomological) relation to exp_R.

Note that Mary can come to know about her experience only through the exercise of her concepts applying to it. This is required by Jackson's argument,

which is about factual / propositional knowledge (knowledge-that as opposed to knowledge-how). Now the materialist reply to Jackson we favor can be stated more explicitly and clearly. The extensions of Mary's two concepts, EXP_R and SDE_R,[11] are numerically identical, i.e. these concepts denote the same property. Assuming that Mary can project her essentially perspectival concept, EXP_R, on to other people's experiences of red, we can then respond to Jackson's second premise in two ways. We can grant it under one reading that takes the novelty involved not as a novelty in facts, but as a novelty in representing facts. But this is harmless for physicalism: Jackson's conclusion does not follow. Or we can read (2)' as claiming that Mary comes to discover a new fact which she did not represent (let alone, know) before. But then the premise is false.

We think that this response to the Knowledge Argument is fully satisfactory from a technical viewpoint, and the diagnosis it embodies is compelling: namely, that the apparent incommensurability between our grasp of what it's like to visually experience red and of what underlies it physically stems from two radically different ways of epistemically accessing the same (physical) phenomenon; it is not indicative of a dualist ontology. There are antecedents to this kind of dual epistemic access. For instance, consider the fact that we have discovered that water is H_2O, that lightening is a certain kind of electrical discharge, that temperature is mean molecular kinetic energy, etc. These are all a posteriori identities, revealed by scientific investigation. In each case, the two concepts flanking the identity sign are radically different in character, though they pick out the same phenomenon.

There are of course very significant differences between the two paradigms, viz., between ordinary scientific and psychophysical identities, and some of these are very important.[12] For instance, it seems plausible to claim that in the case of scientific identities there are no mysteries in explaining how certain facts, say, about water, turn out to be facts about H_2O, whereas the mystery seems to be persist with full force in the case of claims about colour experience. We agree. But we claim that the reason for this stems from a peculiar feature of our sensory and introspective concepts, the explanation of which points to the right sort of HOR account.

6 Introspection and Phenomenal Concepts

The concept EXP_R should be distinguished from Mary's concept RED, which denotes the redness possessed by physical objects. Before her release, Mary certainly had a concept that she could express with the English word 'red', and which she probably associated with her scientific conception of redness, SD_R; but it was not the same concept that lay people with normal colour vision express when they use 'red'. Mary's newly acquired RED applies to physical objects, but her EXP_R applies to her experiences of red, i.e. to exp_R. How does RED differ from SD_R?

It seems clear that RED is directly and immediately acquired from experiences of redness, whereas SD_R is not. For the acquisition of the perspectival concept (EXP_R), not only having exp_R but also having RED is necessary. So we have a new set of dependencies:

redness <— exp_R <— RED <— EXP_R,

While RED denotes (is about) redness, EXP_R denotes (is about) exp_R. However, the way each of these acquires its denotation is quite different. RED is a simple and atomic concept directly acquired from exp_R without the mediation of any other concept. Even though redness is a complex physical property, its analog representation in our experience is simple, i.e. it does not reflect the physical/structural complexity of what it represents; on the contrary, it represents redness as a simple property. RED is acquired from such a basis. So it preserves the semantic simplicity of its analog counterpart. Let us expand on these remarks a bit.

Sensory experiences are supposed to track changes in the environment. In this they are (analog) representations whose primary job is to make available to their HOSTS temporally indexed information about their environment. This is very important: sensations are responses, responses to environmental changes. As such their informational value is restricted within a time frame sufficient for the organism to act back on the environment effectively on the basis of this information. In short, sensory representations are stimulus-driven. We will call this vertical information processing.

By contrast, thinking and reasoning (like daydreaming and imagining) are horizontal forms of information processing. By this we mean that they can, and pretty frequently do, occur in the absence of a direct or immediate causal relation with the things being thought or reasoned about. This is perhaps the most important hallmark of human intentionality. We harbor representational processes

that are not directly prompted by what those processes are about. But thought and thinking require concepts.

Although all concepts can be causally decoupled from their referent and thereby implicated in horizontal processes, many of them can also be used for vertical informational purposes, i.e. uses such that their tokenings indicate the instantiation of the property they denote. In this (extended) vertical process, experience is the necessary intermediary. In fact, perception, unlike sensation or mere experience, is the vertical process whereby objects of sensation are cognized and recognized, i.e. categorized or sorted under concepts. For most observational concepts, this takes the form of recovering the information already (mostly) in the sensory array by computational processes that eventuate in the tokening of the concept. We regard this process mainly as one of information extraction by digitalization / abstraction from a rich array of information present in analog form in the experience.[13] The mechanism underlying the formation of primitive sensory concepts and their vertical deployment is probably hard-wired in organisms like us.

So the relation between experience and thought, in particular, between the sensory representation of redness, i.e. exp_R, and RED comes down to this. Tokenings of exp_R (analog) are normally the result of vertical processing. They are normally under the nomic control of redness: they are stimulus-driven, whereas tokenings of RED (digital) may be causally independent of this property. Unlike exp_R, RED is the kind of cognitive state or structure that is capable of involvement in horizontal processing. For our purposes we can treat concepts, following CRTM, as terms of a

language of thought realized in the brain of sufficiently sophisticated cognitive organisms.[14]

All phenomenal concepts, like RED, are concepts acquired[15] directly from the representational content of experience. We believe that experiences of so-called subjective "secondary qualities"[16] are simple and semantically primitive representations of complex physical properties whose instantiations directly prompt their tokening under appropriate circumstances. As such, phenomenal concepts are direct classificatory responses to physical stimuli as represented in the experience. They need not be predicative, they can be demonstrative and as fine-grained as our discriminative capacities can allow with respect to the relevant dimensions of the physical property being detected. What needs emphasis here is that our phenomenal concepts are concepts that represent physical (external as well as bodily) determinables as represented by our experiences. They do not apply to our experiences.

As we have mentioned, our concept formation and application mechanisms are built in such a way that phenomenal concepts are those that are acquired or applied without the mediacy of other concepts. In this, the informational relation between the exp_R and RED is brute and unanalyzable — semantically or epistemically. Because the type of information in the experience to which RED is a direct classificatory response is encoded by A simple and primitive type of informational response to complex physical properties, RED represents the physical property it does as a primitive.[17]

That there could, or even should, be such concepts seems obvious given a rough outline of our cognitive and sensory architecture — and how these systems interface. RED, very much like exp_R, picks out a certain sort of complex physical property (or properties) directly, without representing its internal structure. This is the moral of a certain sort of vertical information processing eventuating in conceptual categorization: that it requires simple, primitive concepts whose tokenings are direct and unmediated, even though what they detect may be complex. We can summarize our discussion of the relation between RED and SD_R by saying that they have radically different causal/functional/conceptual roles: their acquisition and application conditions are radically different. One requires actually occupying certain sort of informational states for its acquisition and application (vertical/classificatory response situations); the other does not, but gains its "cash value" primarily by its role it plays in horizontal cognitive processes. Its acquisition is not direct but heavily mediated by various other concepts, including sensory ones.

EXP_R, in contrast, is not simple in the way RED is — as long as it requires the possession of RED. We see EXP_R as a concept whose semantic content is [experiencing red], so in this it is like EXPERIENCING RED. Still, the complexity involved here is special, quite different from say, the concept BACHELOR. It is by appeal to this special character of EXP_R that we hope to ultimately explain why psychophysical identities seem so puzzling and mysterious even if we were independently convinced that physicalism is true.[18] In order to capture this special character, we need to apply the account of the sensory/perceptual processes just described to the acquisition of EXP_R.

Given the multiplicity and richness of vertical information entry in organisms like us, it is clear that which particular sensory channels are activated in particular cases (and through which parameters of each channel) is itself a source of information. Even though each particular set of sensory channels in each modality is supposed to be stable and, as such, not an information generator, at a higher level in the cognitive hierarchy, the variability in the activation of any particular set among all is an information generator — if we have introspective ways of monitoring the activity in the various channels.

We speculate that we do have this capacity. We propose that at least in humans there are introspective mechanisms dedicated to monitor the avenues of vertical information entry eventuating in categorization of the information present in the experience. If this is right, we can model it in much the same way as the sensory systems eventuating in the perception of distal (secondary) properties. As we said, perception requires categorization, which is a minimal conceptual capacity. In this we regard introspection of our experiences[19] as itself being sensory-cum-perceptual. So we propose that EXP_R is acquired partly as a result of a sensory-like introspective mechanism which monitors and detects in a simple/primitive way which vertical information avenues are active, and eventuates in introspective categorization in conjunction with the conceptual deliverances of the sensory channels proper. So, this mechanism takes in the conceptual deliverances of sensory channels as input along with a sort of primitive vertical detection of the particular sensory channel being activated, and delivers as output an internal

perception, i.e. conceptual categorization, like the tokening of EXP_R as a vertical response to exp_R.

One important aspect of this sort of introspective vertical processing is its sensitivity to the temporal window or duration of the activation of the channels. Probably this is one of the major sources underlying the tradition that regards introspection as a sort of internal sensing or monitoring. But again it is worth emphasizing that this monitoring is hard-wired to eventuate in a minimal sort of categorization in the sense of registering information about channel conditions along with the sensory concept used to vertically sort the analog information in the sensory experience itself. The main reason why introspection seems to be transparent, i.e. why the properties we encounter when we introspect our experiences seem all to be the properties that our experiences detect rather than exhibit, is that introspection eventuates in the tokening of a composite concept whose more substantive constituent is borrowed from the sensory categorization processing itself while active. We are not sure how to proceed further in our speculation at this point, but what is essential for our purposes is that introspection mechanisms never work independently of the outputs of the proper sensory channels. What we need is a mechanism that would reflect this nomological necessity. The contribution of introspection is in fact exhausted by the dedicated detection of the simple extra information generated by the activation of particular vertical information avenues. Its output, then, is a tight organic integration of this extra information with the outputs of the proper sensory channels, which are always about the objective features of the environment (external and bodily).

Note that on our account introspection is impossible without the relevant concepts of both sorts, sensory proper and the quasi-demonstrative/indexical concepts belonging to introspection proper. But both the acquisition and the vertical deployment of these concepts are radically perspectival, first-person. In properly working organisms with the relevant sort of sensory/perceptual/cognitive machinery intact, the acquisition and the processing profile of such concepts will be host-unique in that such a profile is hostage to the actual workings of vertical information systems.

Note also that this proposal about our introspective awareness of phenomenal states is a synthesis of HOP and HOT. We regard introspection of phenomenal states as a process of sensation-cum-perception. The sensation part yields HOP; but introspection produces conceptually articulated cognitive states, hence we have HOT.

7 Conclusion

If our speculations are even approximately true, Mary can acquire and deploy the introspective concept of experiencing red, EXP_R, only after she sees red for the first time. Furthermore, given the concept formation mechanisms outlined above, her EXP_R would directly pick out the same event as her scientific conception of red experience, SDE_R. But now think of the enormous disparity between the acquisition and the deployment of these two sorts of concepts.

The story we have told is not wild metaphysics. It is the kind of story that is in principle empirically testable, as well as constructible in an AI system. Since we

are convinced of the truth of physicalism, we suspect that further empirical research into human introspective capacities will bear us out. It should be obvious that if we are right in our diagnosis, AI research on the construction of robots with phenomenal consciousness has a clear path to follow: Build systems with a sensory / cognitive / introspective architecture which is in the spirit of our proposal. If we are right, robots need not be spared phenomenal consciousness. In fact, if constructed in the way we suggest, they just might be as curious as we are about how it is that this is what seeing redness could possibly be.[20]

ACKNOWLEDGEMENTS: We would like to thank Varol Akman, John Kulvicki, and Philip Robbins for their helpful comments on an earlier version of this paper.

References

Aydede, Murat, 1997, Language of thought: the connectionist contribution. Mind and Machines, 7: 57–101.

Aydede, Murat, 1998, Language of thought hypothesis: state of the art. Draft, The University of Chicago, available at http://humanities.uchicago.edu/faculty/aydede/LOTH.SEP.html.

Aydede, Murat and Güzeldere, Güven (in prep.) Introspection and phenomenal properties.

Block, Ned, 1980, Troubles with functionalism. In N. Block (editor) Readings in Philosophy of Psychology, Vol. 1 (Cambridge, MA: Harvard University Press), pp. 269–305.

Chalmers, David, 1996, The Conscious Mind (Oxford, UK: Oxford University Press).

Churchland, Paul M., 1990, Knowing qualia: a reply to Jackson. In P. M. Churchland, A Neurocomputational Perspective: The Nature of Mind and the Structure of Science (Cambridge, MA: MIT Press).

Descartes, René, 1664, Treatise on Man. In The Philosophical Writings of Descartes: Volume I. (Cambridge, UK: Cambridge University Press).

Field, Hartry H., 1978, Mental representation. In S.P. Stich and T.A. Warfield (eds.) Mental Representation: A Reader (Oxford, UK: Basil Blackwell, 1994), pp. 34–77.

Fodor, Jerry A., 1987, Psychosemantics: The Problem of Meaning in the Philosophy of Mind (Cambridge, Massachusetts: MIT Press).

Fodor, Jerry A., 1990, A Theory of content (parts I & II). In J. A. Fodor, A Theory of Content and Other Essays, (Cambridge, Massachusetts: MIT Press).

Fodor, Jerry A., 1991, Replies (Ch. 15). In B. Loewer and G. Rey (eds.) Meaning in Mind: Fodor and his Critics (Oxford, UK: Basil Blackwell), pp. 255–318.

Fodor, Jerry A. and Pylyshyn, Z. W., 1988, Connectionism and cognitive architecture: a critical analysis. In S. Pinker and J. Mehler (eds.) Connections and Symbols (Cambridge, Massachusetts: MIT Press, A Cognition Special Issue), pp. 3–71.

Güzeldere, Güven, 1995, Is consciousness the perception of what passes in one's own mind? In Thomas Metzinger (ed.) Conscious Experience (Paderborn: Schöningh), pp. 335–357.

Güzeldere, Güven, 1997. The Many Faces of Consciousness: A Field Guide. In Ned Block, Owen Flanagan and Güven Güzeldere (eds.) The Nature of Consciousness: Philosophical Debates, (Cambridge, Massachusetts: MIT Press), pp. 1–67.

Harnad, Stephen, 1990, The symbol grounding problem. Physica, 42: 335–346.

Haugeland, John, 1985, Artificial Intelligence: The Very Idea, (Cambridge, Massachusetts: MIT Press).

Hobbes, Thomas, 1651, Leviathan. Reprinted in 1986 (Middlesex: Penguin Books).

Jackson, Frank, 1982, Epiphenomenal qualia. Philosophical Quarterly, 32: 127–136.

Jackson, Frank, 1986, What mary didn't know. Journal of Philosophy, 83(5): 291–295.

Jackson, Frank, 1994, Armchair metaphysics. In John O'Leary Hawthorne and Michaelis Michael (eds.) Philosophy in Mind, (Dordrecht: Kluwer), pp. 23–42.

Jackson, Frank, 1998, Postscript on qualia. In Frank Jackson, Mind, Method and Conditionals, (London & New York: Routledge), pp. 76–79.

Kulvicki, John, 1999, Ph.D. Dissertation in progress. The University of Chicago, Philosophy Department.

La Mettrie, Julien Offray, 1748, Man, A Machine. Reprinted in 1994 (Indianapolis: Hackett Publishing Co.).

Levine, Joseph, 1993, On leaving out what it's like. In Martin Davis and Glyn W. Humphreys (eds.) Consciousness, (Oxford, UK: Basil Blackwell), pp. 121–136.

Levine, Joseph, 1998, Conceivability and the metaphysics of mind. Noûs, 32(4): 449–480.

Loar, Brian, 1997, Phenomenal states. In Ned Block, Owen Flanagan and Güven Güzeldere (eds.) The Nature of Consciousness: Philosophical Debates, (Cambridge, Massachusetts: MIT Press), pp. 597–616.

Lycan, William G., 1987, Consciousness (Cambridge, MA: MIT Press).

Lycan, W., 1996, Consciousness and Experience (Cambridge, MA: MIT Press).

Margolis, Eric, 1998, How to acquire a concept. Mind and Language, 13(3): 347–369.

McCarthy, John, 1999, Making robots conscious of their mental states. Draft, revised version of a paper with the same title that first appeared in S. Muggleton (ed.) Machine Intelligence 15, 1995 (Oxford, UK: Oxford University Press).

Rey, Georges, 1992, Sensational sentences switched. Philosophical Studies, 67: 73–103.

Rey, Georges, 1993, Sensational sentences. In M. Davies and G. Humphrey (eds.) Consciousness, (Oxford, UK: Basil Blackwell), pp. 240–257.

Schiffer, Stephen, 1981, Truth and the theory of content. In H. Parret and J. Bouvaresse (eds.) Meaning and Understanding, (Berlin: Walter de Gruyter).

Searle, John R., 1980, Minds, brains, and programs. Behavioural and Brain Sciences III(3): 417–424.

Simon, Herbert, 1969, The Sciences of the Artificial. Reprinted in 1988 (Cambridge, Massachusetts: MIT Press).

Tye, Michael, 1995, Ten Problems of Consciousness: A Representational Theory of the Phenomenal Mind (Cambridge, MA: MIT Press).

Van Gulick, Robert, 1993, Understanding the phenomenal mind: are we all just armadillos? In M. Davies and G. Humphrey (eds.) Consciousness, (Oxford, UK: Basil Blackwell), pp. 137–154.

Notes:

[1] It is interesting to note here that a contemporary of Descartes, Julien de La Mettrie, used Descartes's work to turn his argument around, and claim that human beings, since they are not unlike all other animals in being continous products of nature, must also be machines, whose behavior can wholly be explained in terms of bodily and brain processes: 'Man is a machine so complicated that it is impossible at first to form a clear idea of it, and, consequently, to describe it. This is why all the investigations the greatest philosophers have made a priori, that is, by wanting to take flight with the wings of the mind, have been in vain. Only a posteriori, by unraveling the soul as one pulls out the guts of the body, can one, I do not say discover with clarity what the nature of man is, but rather attain the highest degree of probability possible on the subject' (La Mettrie 1748: 30).

[2] But see Rey (1992, 1993) for an attempt to extend CRTM in this direction.

[3] This is to convey the basic idea that each type of attitude (e.g. believing) is realized by a unique computational relation (e.g. being inside the computationally defined B-Box — "Belief-Box") . So the mapping from attitudes A into computational relations R is meant to be injective.

[4] Naturalizing intentionality (Fodor's second problem), as opposed to intelligence physical processes (third problem), may be independent of a computational framework. In fact, as the formulation of CRTM makes clear, CRTM assumes in (B) that the internal language already has a semantics, so in this sense it assumes that there is a naturalistic story to be told about how this language acquires its semantics in the first place. (This is highlighted by Searle's famous Chinese Room thought-experiment, see Searle 1980; in the AI community and among psychologists, this problem is discussed under the label 'the symbol grounding problem', see Harnad 1990.) But it is natural to expect that this story would require the resources of computationalism. So to this extent, CRTM may not itself provide a complete solution to the second problem, but would nevertheless provide some crucial help in obtaining the ultimate solution. For an attempt to naturalize intentionality along these lines, see Fodor 1990. For a longer and more detailed survey of CRTM and its place in the history of contemporary philosophy of mind, see Aydede (1997, 1998).

[5] For a review of philosophical problems of consciousness, see Güzeldere (1997).

[6] Jackson has recently gave up his anti-physicalist position, see the postscript on qualia in his (1998).

[7] The logical possibility of such a scenario is defended by Block (1980), Jackson (1994), Chalmers (1996).

[8] For a more detailed treatment of the distinction within HOR theories and a critique of the HOP accounts, see Güzeldere (1995).

[9] See, for instance, Lycan (1987, 1996), Churchland (1990), van Gulick (1993), Tye (1995), and Loar (1997).

[10] When we denote concepts, following the standard practice, we will use the small cap font. We take concepts to be mental representations realized in the brain.

[11] The concept SDE_R stands in lieu of the conception of a scientific description of experiencing red; i.e. the kind of psychological state Mary expresses when she uses the English equivalent of 'sde_R'. In denoting the concept of sde_R, as noted we are using small cap 'SDE_R'. SDE_R is a psychological representation realized in her brain whose reference is also realized in her brain upon actually experiencing red for the first time.

[12] We are not committed to the view that mental types are reducible to physical types. One can reject this sort of very strong type-type identity view in favor of a weaker metaphysical supervenience thesis: that the mental strictly metaphysically supervenes on the physical, which amounts to a thesis of token-token identity.

[13] The contrast between sensation and perception is sometimes commented on by philosophers in terms of a distinction between seeing and seeing as (hearing and hearing as, etc.). Anyone with normal vision can see an aardvark, but if they have no idea of what aardvarks are (not having the concept of one), they cannot see it as

an aardvark. In 'S sees x', 'x' occurs transparently (could be replaced by any coreferring expression without changing its truth value), but the occurrence of 'F' in 'S sees x as F' is opaque, reflecting the fact that the truth-value of the statement depends on whether S has the concept expressed by 'F' and applies it to x as a consequence of standing to x in the seeing relation.

[14] Let us also make the assumption (mostly heuristic for the purposes of this paper) that there is a central cognitive system where horizontal processing at the conceptual level occurs, and in which confirmation, belief fixation, and decision making are more or less holistic, and are sensitive to the global pragmatic and epistemic properties of the whole system (conservatism, simplicity, and the like).

[15] Our use of the term 'acquired' as applied to concepts is intended to be neutral with respect to the nativism/empiricism debate. So, if you think that the right notion is "triggered" rather than "learned", that is perfectly fine with us. In fact, we are sympathetic. Our view, on this matter, is very much like the view of Margolis (1998). But even for empiricists, the right notion for phenomenal concepts would be "triggering" rather than "learning", since for them such concepts are the semantic primitives from which all other concepts are constructed.

[16] Like being red, sweet, sour, warm, etc. as opposed to being round, rough, moving, which are sometimes called, "primary qualities" that are regared objective.

[17] There are, however, complications that we want to ignore in this paper. Strictly speaking, sensory representations of non-unique hues, like purple and orange,

represent them as having some complexity relative to unique hues. So what we need to maintain is that such sensory representations of secondary qualities may represent them as consisting of something complex without telling what that "something" (i.e. surface spectral reflectances) is. Kulvicki (1999) works out this line of thought convincingly in quite elaborate and original ways in the context of discussing intuitions underlying spectrum inversion.

[18] This puzzling feature of psychophysical identities (or, in general how to genuinely understand how physicalism can be true even if we are convinced that it is true) is sometimes pointed to as evidence of an explanatory gap between the physical and the phenomenal, even though the gap may not be ontological. See Levine (1993, 1998).

[19] We want to restrict our analysis of introspection here to the introspection of experiences. We are open to the idea that there may be a multiplicity of introspective mechanisms of different sorts for a variety of different mental phenomena.

[20] For a more elaborate discussion of our diagnosis and proposal, see Aydede and Güzeldere (in prep.).