Minds & Machines
DOI 10.1007/s11023-012-9285-z

How the Problem of Consciousness Could Emerge in Robots

Bernard Molyneux

Received: 17 July 2011 / Accepted: 29 June 2012

Abstract I show how a robot with what looks like a hard problem of consciousness might emerge from the earnest attempt to make a robot that is smart and self-reflective. This problem arises independently of any assumption to the effect that the robot is conscious, but deserves to be thought of as related to the human problem in virtue of the fact that (1) the problem is one the robot encounters when it tries to naturalistically reduce its own subjective states (2) it seems incredibly difficult from the robot's own naturalist perspective and, most importantly, (3) it invites the robot to engage in the exact same metaphysical responses as humans offer to the problem of consciousness. Despite the fact that it invites the robot to consider extravagant metaphysical solutions, the problem I explore is purely algorithmic. The robot cannot complete its naturalist physicalist reduction as a matter of algorithmic fact, whether or not the naturalist physicalist reduction would be correct as a matter of metaphysical fact. It is hoped that by reproducing the familiar seeming problem in an artificial context, a greater understanding of the human problem of consciousness can be achieved.

Keywords Hard problem · Consciousness · Strong AI · Artificial consciousness · Reduction · Identification regress · Explanatory gap · Identity · Mysterian

Introduction

Nagel (1974) famously remarked that the problem of consciousness makes the mind body problem interesting, while also making it intractable. Instead of attempting to solve what appears unsolvable, an alternative reaction is to investigate why the problem seems so hard. In this way, Minsky (1965) hoped, we might at least explain

B. Molyneux (✉)
Department of Philosophy, 1240 Social Sciences and Humanities, University of California, Davis, One Shields Avenue, Davis, CA 95616, USA
e-mail: molyneux@ucdavis.edu

Published online: 17 July 2012

why we are confused. Since a good way to explain something is often to build it, a good way to understand our confusion may be to build a robot that thinks the way we do. I will explore the lessons of such an endeavor in what follows. I hope to show how, by attempting to build a smart self-reflective machine with intelligence comparable to our own, a robot with its own hard problem, one that resembles the problem of consciousness, may emerge.

The strategy is to show that when the robot tries to identify something that it encounters in its third-person observations with something from its 'subjective' 'inner' life, it either falls into algorithmic regress or starts to behave like humans; i.e., by either reacting 'dualistically' and adding its inner states and properties as new and irreducible features of its ontology; by becoming an eliminativist and denying that its inner states and properties are really there at all; or by adopting that unsatisfactory, 'forever unfinished' version of physicalism wherein, though it insists these inner states are identical to *something* physical, it cannot find a plausible candidate. I will refer to this as the robot's 'hard problem'.

If we assumed that the robot had a conscious inner life and then showed that, just like ours, it was difficult to reduce, we would only expose our existing confusion over the phenomenon of consciousness, wherever we imagine it to be. To get a better explanatory grip on consciousness, rather, we should show that having assumed *nothing* about the presence or absence of consciousness in the robot, its inner states end up being procedurally hard (and perhaps impossible) to reduce *nonetheless,* from its point of view. In this way, we show how one of the principal properties associated with a state's being conscious—i.e., its seeming irreducibility from the viewpoint of its subject—naturally comes to attend the inner states of the robot, arising non-magically from ordinary algorithmic procedures.

Accordingly, we must formulate a cold, non-phenomenal, purely informational, notion of 'subjective experience' so that we can imagine our robot to have a subjective inner life without imagining it to be conscious. We do this in the third section, where we distinguish the robot's private, subjective phenomena from the regular phenomena it finds in its third person space. We shall then, in the fourth section, make the robot into a naturalist reductionist, for if we are to see how the robot tries and fails to identify its subjective states with something objective, we need to program it to try. In the fifth, we show how, from all this, the robot's hard reduction problem emerges. And in the sixth, we clarify our conclusion, dealing with possible objections and misunderstandings. Before all that, we need to understand something about how the machine identifies things, and how it reasons quite generally. We attend to that in the next (second) section, where we introduce some rules and idealizations to ensure the robot reasons well.

How the Robot Thinks

Let's suppose we build a robot, M, that can investigate and reason well about 'his'[1] environment. Though terms like 'reasoning' have mentalistic connotations we wish to avoid supposing that M is conscious. Rather, we will imagine M's inferential

[1] We give M 'his' own arbitrary gendered pronoun for readability.

process to be an ordinary computational procedure in which representations are systematically manipulated to produce new veridical representations from old. So, for example, M might proceed from symbolic representations of *All men are mortal* and *Socrates is a man* to one of *Socrates is mortal* simply by performing appropriate formal manipulations.[2] Other seemingly-mental terms should, throughout what follows, be given a similarly deflationary reading. P *seems* to be true, for example, when M's sensory inputs make P-information available to the system as a whole, influencing behavior and verbal report, whether or not there is 'something it is like' for M when this happens. And for M to *believe*, *want* or *experience* something, he merely needs to process and take action *as if* he believed, desired or experienced it. Consciousness is never assumed.

Leibniz's Law

The principal rule of good reasoning we impose upon M is:

> *Agreement with Leibniz's Law*: M's identifications are always in accordance with Leibniz's Law, a.k.a. the indiscernibility of identicals.[3] I.e., if M holds that A = B then, for every property P, M holds that A instantiates P if and only if M holds that B does.

Leibniz's Law, in plainer language, is simply the claim that everything has the same properties as itself. I will sometimes talk about thinkers *violating* Leibniz's Law, but I do *not* mean that the thinkers have different properties from themselves, for that is impossible. I mean only that they fail to observe Leibniz's Law *in their reasoning*. Since M would be a poor reasoner if he violated Leibniz's Law, we stipulate that he does not.

Other Assumptions

In a similar spirit, we make the following simplifying assumption: That if M seeks to identify X with Y and if M is positioned to observe or infer that X has some property P that might possibly impede identification, then M does so. For example, suppose that M seeks to identify the apple in his right hand with the orange in his left, and is positioned to observe that the former is green. According to our simplifying assumption, M makes the observation. Of course, when he also observes that the orange isn't green (and that it's in a different hand!) this blocks him from identifying the apple and the orange, just as it would a human.

Because it can block prospective identifications, our simplifying assumption does not simplify things for M. On the contrary, it consistently makes it harder to identify things. This difficulty, however, is the result of good practice. It arises from building M to be a diligent investigator. And however it complicates things for M, it

[2] We can grant, if intentional content requires consciousness, that these tokens have content only from our human perspective. I.e. that they are merely truth evaluable, or otherwise semantically evaluable, *by us*.

[3] *Not* the converse (and more controversial) identity of indiscernibles, for which the term 'Leibniz's Law' is also sometimes used.

simplifies things for us. For instead of imagining that M identifies apples and oranges then reverses his identification later, when he notices their colors, we just assume that M notices their colors at the beginning.

We help ourselves to other assumptions of a similar sort: That M does not forget his own definitions; that he does not make mistakes when deducing; that he does not confuse one thing for another, and so on. They can be grouped together under the assumption that M is *epistemically diligent*. The hard problem that follows relies, in part, upon these assumptions. A robot of which they only approximately held would have its own, slightly more messy, hard problem. For though it might mistakenly solve its hard problem by neglecting identification-blocking properties, or by otherwise making mistakes, it would, if it were diligent *enough*, renege on its solution later. If, on the other hand, it *wasn't* diligent enough, and never realized what it overlooked, mistook, etc., then it would fall beneath the class of machines that are epistemically responsible enough to have their own hard problem.[4]

Objective and Subjective Phenomena

Objective Phenomena

Shortly, we will be talking a lot about *subjective* phenomena. It is often thought, though, that we live in a world which, in principle, can be described entirely from an *objective*, third person, perspective. If that's right, then each subjective phenomenon is some objective physical event, for example a brain process. It would not be correct, in that case, to split the world into subjective and objective phenomena, as if they were distinct realms, since all subjective phenomena are just objective phenomena observed from a particular point of view. The phrases 'subjective phenomena' and 'objective phenomena', then, in what follows, should not (necessarily) be thought of as designating classes of events that are distinct and independent, but as shorthand for, respectively, 'phenomena presented subjectively' (or 'considered from a particular point of view' or 'via a subjective concept') and 'phenomena presented objectively' (likewise). So when I consider the conscious experience of eating an apple from the first person, through introspection, it presents itself as a subjective phenomenon. Assuming that same experience is a brain process that anyone can watch on a monitor then, considered as such, the experience is a perfectly objective phenomenon.

All of the following count as core cases of what we mean by 'objective phenomena':

> rooms; electrical current; sticks of wood; potential difference; holes in the earth; rocks; quarks; the property of being a cat; the process of norepinephrine re-uptake; the state of being solid; heat; neural events; signal transmissions;

[4] Obviously, the conclusion that an epistemically diligent M-machine cannot complete the task entails the conclusion that if an M-machine completes the task, it was not epistemically diligent. An M-machine could therefore have a belief in the identity, even a *correct* belief, but, assuming that such things require epistemic diligence, not *justified* belief or *knowledge*.

stamp collecting; solubility; sunrises; songs; monetary transactions; aunts and uncles; leaps; greetings; emancipation; centers of gravity; smiles; etc.

An exhaustive list would include everything that one could learn about from the third person, which is why the sample is so taxonomically liberal, cutting across states, objects, events, properties and processes. Abstracta, as well as concreta, count as objective. The set of all bloodhounds, the number three and the tiger-in-general are all to count as objective phenomena.

Since all of the following are characterized functionally, teleologically or informationally, they too should count as objective:

the inclination to scratch; the disposition to aversive behavior; the tendency to cause yelps; the property of having been selected to correlate with tissue trauma; the property of being targeted by a higher order intentional state; representations of redness; signals communicating tissue damage; etc.

The phenomena listed above are among the kinds of thing associated with functional characterizations of conscious states. A conscious itch on the toe may, for some theorists, be characterized via a network of functional dispositions that includes the inclination to scratch. The pain of an animal may be identified with a functional kind characterized by, *inter alia*, the disposition to aversive behavior and the tendency to cause yelps. And so on. Hence, if our conscious states are functional, teleological or otherwise supervenient states, they will be found (under that guise) on the exhaustive list of all objective phenomena.

Reduction as *Identification* By counting supervenient phenomena among the objective in this way we are able to univocally characterize *reduction* as *identification*; for there is no difference between saying that the Fs supervene upon the Gs and saying that the Fs are *identical* with things *that* supervene upon the Gs. Whatever subjective states are, then, one *reduces* them to the objective by *identifying* them with some (quite likely supervenient) member of the list of all objective phenomena. When we introduce M's subjective phenomena, the same will hold true. In order to *reduce* his subjective phenomena M must *identify* them with something on the list of everything objective.

The Order of Phenomena It turns out to be useful, in the forthcoming demonstration, to be able to keep track of the order of a phenomenon—i.e., to explicitly distinguish properties of objects from properties of properties, and those from properties of properties of properties, and so on. We use superscripts to this end. The superscript on 'X^1', then, indicates a phenomenon at the first order;—an ordinary property—whereas the one on 'X^2' indicates a property of an ordinary property, and so on. We can naturally let the superscript '0' indicate non-properties—i.e., ordinary objects like sticks, stones and fire trucks.[5]

We use the symbols 'X' and 'Y' to represent arbitrary objective phenomena, so that anything proved of X or Y is true of any objective phenomenon. We will tend to use 'X' a lot, so the fact that we used it for both X^1 and X^2 in the last paragraph

[5] Since nth order properties always apply to n-1th order phenomena.

should *not* be taken to indicate that we mean the same property at different orders. X^1 and X^2 may well be completely different properties. Whenever the superscripts distinguish one 'X' from another in this way, there is no need to use another letter, and so we will tend to use 'Y' only when necessary.

Subjective Phenomena

So what are subjective phenomena? We can imagine that M first discovers them during a lab accident when water is spilled onto his exposed circuits. Malfunctions result, which cause M to report a red fire truck, moving left to right, sirens blaring. Upon learning that there is no fire truck, M realizes that he is merely in the state he *would* be in if a red fire truck were passing by. This state—the one that would normally constitute the veridical detection of a fire truck—is the first subjective phenomenon discovered by M.

We must not assume, of course, that M's illusion is conscious. Fortunately, nothing about the scenario forces us to. For the illusion depends only upon M's inability to distinguish his current situation from the one where the fire truck really is there. It therefore depends only upon M's functional inability to make certain discriminations, whether or not it is consciously *like something* for M to fail in this way.

If the state is not assumed to be conscious, how is it subjective? It is subjective because M has a privileged informational perspective on whether he is in it. Whereas the question of whether a fire truck is *really* going by is one that third person observers can answer as easily as M, the same is not true of whether M is detecting one. We can infer that M is detecting a fire-truck from its actually driving by, on the assumption that M's sensors are in working order. And if we access and decipher his inner data states, we might laboriously conclude that he detects one. However, it is always quicker to simply ask M, and this alone hints that, however, good our access, M has better access to the states we are after. Indeed, it is plausible that M's earnest reports of a fire truck *count* him as detecting one—that such earnest reports therefore *cannot be wrong*—a possibility we will discuss more later. For now, though, we will say only that M has his own special (good or bad) way of determining whether or not he is in these states, and it is one that we cannot share. It is this that counts them as subjective.

Subjective Correspondents M's illusion, despite not being assumed conscious, can nonetheless prompt him to have thoughts similar to those that might occur to a philosopher of mind. Imagine, for example, M reasoning as follows:

> "I was wrong," says M "that there was a fire truck nearby. However, it was true that there *seemed* to be one. So the truth of the seems-statement did not depend on the fire truck itself, for there was none, but on my own inner states. It appears, then, that there exists a state of my being, which I will call 'the *subjective correspondent* of the fire truck', which—when and only when I am in it—makes it true that there seems to be a fire truck. Generally, where O is an object (or other non-property) the subjective correspondent of O is that state of my being that makes it true that there seems to be an O."

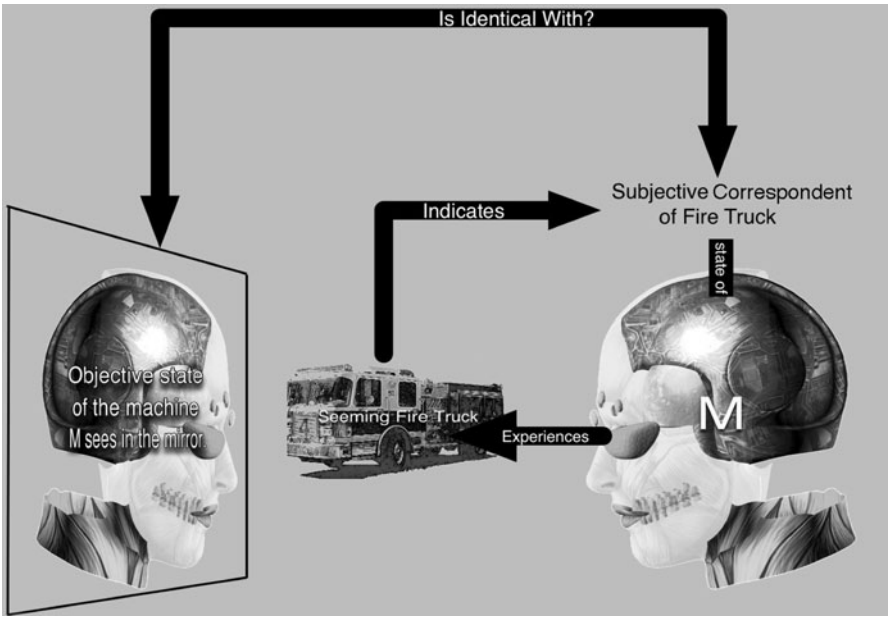

Fig. 1 Setup for M's hard problem. Since the fire truck is illusory, the statement 'there seems to be a fire truck' is not made true by it, but by some state of M. Call that state the 'subjective correspondent' of the fire truck. M comes to wonder which objective phenomenon this state is identical to. Perhaps it is some physical state of the machine M sees in the mirror

We shall follow M in supposing that if (veridical and illusory) detections of fire trucks exist, then some M-state (however, 'gerrymandered', 'non-natural' or multiply-realized) must account for all such detections.[6] M goes on to observe (Fig. 1):

> "The fire truck, moreover, seemed to be *red*. In seeming so, it seemed the way that real and illusory tomatoes, strawberries and communist flags seem. This sameness cannot be in the things themselves, since it is present even when I experience an illusion. I must conclude that it is a feature of my subjective states, since only they are common to both veridical and illusory experiences. When the fire truck seems to be red, therefore, that must consist in the subjective correspondent of the fire truck having a certain property—the one that also modifies the subjective correspondents of tomatoes and strawberries.

[6] Strictly speaking, nothing in the following requires us to agree with M's arguments. To derive M's hard problem, we do not need something *real* that M cannot reduce: We only need something that M *believes* in.* We could therefore rewrite the following so as to put the existence of a single subjective correspondent into escrow.

* Of course, if the subjective correspondent lacks existence then that explains why M cannot find any objective phenomenon with which to correctly identify it. However, that would be the *uninteresting* explanation. Where Sherlock Holmes seems the kind of thing that *would* be reducible if he existed, conscious experience does not. So if subjective correspondents are to be more like conscious experiences than Sherlock Holmes, we need to show them to be the kinds of things that would seem irreducible on the assumption they exist.

> It is reasonable to dub it 'the subjective correspondent of red'. Generally, then, where P is a property, I define *the subjective correspondent of P* as that property that the subjective correspondent of O bears if and only if O seems to have P."

To facilitate rigorous reasoning in what follows, we can capture M's observations symbolically. Specifically, we can let 'SX^i' designate the subjective correspondent of X^i so that, for example, if X^0 is the fire truck, then SX^0 is the subjective correspondent of it. We use '0' here because the fire truck is an object. For properties, we will use the corresponding higher numbers, so (e.g.,) 'SX^2' refers to the subjective correspondent of the second-order objective property X^2.

M, then, has defined his terms such that PI holds:

> PI: A phenomenon counts as having the subjective property SX^{i+1} iff (a) the phenomenon is some SX^i where (b) X^i seems to have X^{i+1}.

For example, the fire truck (X^0) seeming red (X^1) is necessary and sufficient for the subjective correspondent of the fire truck (SX^0) to count as having, as one of its properties, the subjective correspondent (SX^1) of the redness. If the apparent redness, moreover, seems to have the property (X^2) of being an advancing color and the property (Y^2) of being different from green then the subjective correspondent of the redness will, in turn, count as having the properties (SX^2) and (SY^2), where these are, respectively, the subjective correspondent of (the redness's) advancing coloredness and the subjective correspondent of (the redness's) differentness from green. And so it goes, the broad idea being that we classify the subjective correspondents by commonalities in the seemings they underwrite.

Let us finish this section by asking whether anything answer to the terms SX^i that M defines? It seems that *something* must, since the 'SX^i's are stipulated to be *whatever* it is that is responsible for the relevant commonality across M's experiences. So even if it turns out that no single 'real', 'genuine' or 'natural' property is always responsible, then the relevant SX^i would merely be the (possibly infinite) disjunction of the many properties that are. Suppose that, for example, X^0 is the fire truck and X^1 is the property of being red, and suppose that, when M's circuits are dry, SX^0 has SX^1 if and only if it has the property J whereas, when M's circuits are wet, SX^0 has SX^1 if and only if it has the completely different property K. Then, rather than say there is no unique property SX^1 that plays the role quite generally, M will instead define SX^1 as the disjunctive property of *either being J and such that M's circuits are dry or being K and such that they are wet*.

Nonetheless, the phenomena SX^i are, strictly speaking, M's posits, not ours, and we need neither affirm nor deny their existence to appreciate what follows. I only claim, in fact, that M's way of carving up the matter seems reasonable and prima facie attractive; so that the problem that (soon) results is not to be blamed on some strange analysis on M's part. I could, in fact, if it came down to it, rewrite the following so that we assumed only M's *belief* in the SX^i, not their actual existence. For M can get into a cognitive muddle using only his own concepts, whether or not they map onto the world. It facilitates the exposition, however, if we follow M in treating the SX^i as real.

M's Philosophical Approach

Until now, we have been discussing only how M conceives of his own states; specifically, how he has noticed that he sometimes cannot tell whether the world is really *thus and so* or whether he is merely in that state he would be in if it were. Apart from our brief idealization of M as a rational machine that reasons well, that's really all that has been said so far, with the details presented semi-formally only so as to facilitate rigorous reasoning in what follows.

We turn now to M's attempts to locate his subjective states in objective reality. We impose three restrictions on how M approaches this problem: That he resists the temptation to pretend that things seem other than they do, just in order to make the problem easier; that he approaches the problem as a naturalist—i.e., that he tries to identify each subjective phenomenon with some objectively detectable state; and, finally, that he is not content to merely wave his hands and assure himself that the phenomena in question are identical to *some or other—unspecified*—physical or objective phenomenon. M always seeks to make *specific* identifications.

The First Restriction

M is resistant to revising his beliefs concerning how things seem with him.[7,8] To illustrate the kind of resistance that M exhibits, recall the case where we spill water on his circuits, causing him to malfunction and report a vision of a fire truck. M may well accept that he has made an error about the fire truck, but he won't accept that he made an error about the fire truck *seeming* to be there. He might accept that there is a malfunction in his primary visual sensors or, if that hypothesis fails, switch to blaming his secondary visual abstractor or, if that fails too, ultimately switch to the hypothesis that he is mis-accessing his own visual systems. But no matter how many amendments of this sort he is forced to make, and no matter how difficult it is to locate the precise malfunction, M will never surrender the claim that, when all is said and done, he *seems* to see a fire truck.

It is reasonable to credit M's earnest reports as being a largely infallible guide to how things seem to M. If we were to suppose that M were conscious, after all, then we would only be granting him something like the same authority, when it comes to his conscious seemings, that we usually grant ourselves. If he is *not* conscious, on the other hand, then we should get to *count* M's verbal reports (or the dispositions to make them) as *constituting* how things seem to M. It is, when all is said and done, hard to say what would be a *better* token of how things seem to the M-system taken

[7] There is an intuitive notion—a possible analog of M's commitment—that we humans cannot be wrong about our own conscious states. Alternatively, M's commitment might correspond to 'taking consciousness seriously'—i.e. to not 'redefin[ing] the phenomenon in need of explanation as something it is not' (Chalmers 1996, x).

[8] The fallibilist about conscious experience might, consistent with her view that we can mistake pain for pressure in the dentist's chair, nonetheless concede that *some* experiences are experiences we can't be wrong about. E.g., I cannot be overruled by a doctor when I think I am in agonizing pain. Similarly, a robot without M's infallibillist streak might nonetheless express certainty concerning the nature of *certain* subjective states, With respect to such states, a hard problem can be derived.

as a whole than how it ultimately reacts. Sure—its sensors might report one thing and, because of a malfunction downstream, its recognizers might report another. But the best determinant of the way things *ultimately* seem to the system, given that there is no consciousness to be right or wrong about, is the way the system behaves; in particular, what it reports. That, at any rate, seems like the most plausible position.[9]

Still, nothing much depends on this. For all that we need in the forthcoming arguments is that M *takes* himself to have an infallible authority over how things seem to him. Even if we think he overestimates himself, we nonetheless stipulate that he has the attitude.[10]

M is equally stubborn when it comes to the corresponding subjective states. Officially:

> *First Restriction*: M will not revise his beliefs about which subjective states he enters into, nor about the subjective properties of his subjective phenomena, once he forms those beliefs on the basis of his seemings.

M's stubbornness ought not to seem too strange. There are few ways M could go wrong, after all, in moving from the seems-statement, about which he maintains his own infallibility, to the equivalent statement concerning subjective states. M is hardly likely to err, for example, when he infers from "There seems to be a red fire truck" that "I am in some state *such that* there seems to be a red fire truck." So M claims an infallibilism about his subjective states that goes hand-in-hand with his infallibilism about how things seem. Again, though, for our purposes, the correctness of M's infallibilist stance is irrelevant. All that matters is that M *believes* he is infallible. We do not need him to be right.

The Second Restriction

M aspires towards a purely objective picture of the universe, such that:

> *Second Restriction*: M thinks that every phenomenon can (in principle, ultimately) be identified with some (supervenient) objective phenomenon.

So M thinks that a complete understanding of heat, light, matter, information, intelligence and any other phenomenon can be constructed in principle from concepts solely acquired from the third-person. This is the restriction that encodes what might reasonably be dubbed M's *naturalism*.

[9] The wrinkle in allowing M such infallible authority is that, in some cases, his behaviors may diverge. E.g., asked to press a button only if he seems to see a red flash, M might, through some processing malfunction, press the button while verbally denying that he saw the flash. Similar cases occur in the clinical literature on brain damaged humans. The right response, I think, is not to treat them as counterexamples to infallibility, but to treat them as cases where the agent herself is fragmented.

[10] Given that it does not matter whether M is right or wrong about his own infallibility, one might wonder why I spend so much time emphasizing the plausibility of M's stance. It is because I wish to forestall complaints to the effect that M's problem arises from his adopting an unreasonable position, or from his being unlike humans. If it is granted, though, that M's position is not weird, and that it resembles our own to the appropriate extent, then, even if it is incorrect, we have enough to press forward.

The Third Restriction

We will further assume that, given X and some list, if M identifies X with some member of the list then he insists on identifying X with some *specific* member. Officially, and to introduce a term of art:

Third Restriction: All M's identifications must be *constructive*.

This restriction is odd since it does not have an obvious human analog. Humans could happily commit to (say) Samuel Clement being identical with *some* person from a list of writers without committing to any one in particular. Note, though, that our physical/objective picture of the universe could never count as finished if some acknowledged phenomenon were merely identified with '*some* unknown physical/objective thing or other', without saying *which*. One can interpret the third restriction, then, as encoding the ideal that M's reductions be *finished*.[11] If M escapes his hard problem by violating this commitment, he does so only by refusing to say which objective phenomenon some particular subjective phenomenon is identical to.

M's Problem

At last we are ready to show how M struggles to identify subjective and objective phenomena while observing his restrictions.

Intuitive Version

We start by noting that, at the outset, there are many things that a given subjective correspondent could be. It could be some state of M's circuit board; or of M's whole system; or it could extend beyond M's bodily boundaries. The subjective correspondents do not come 'ready identified' with anything objective. (We will assume, moreover, that M's programmer does not preprogram any of the relevant identities. This assumption will be discussed later.)

For *reductio*, let's assume that M succeeds in making some subjective/objective identification. It follows that, at some point in time, he makes the *first ever* identification of this sort. The first subjective phenomenon, moreover, is *ipso facto* the highest ordered subjective phenomenon to get identified with anything objective (unless there are several phenomena that were jointly the first to be identified, in which case one of them is the highest ordered, and that's the one we are interested in).

To work with a concrete example, let's suppose that this (joint) first, (joint) highest-ordered subjective phenomenon is the subjective correspondent of a green afterimage caused by adaptation in M's photoreceptors after a flash of light. He identifies it, let's suppose, with some state C of his circuits. But thanks to M's diligence in considering potential identification-blockers, M must realize that the subjective correspondent of the afterimage has, as a property, the subjective correspondent of green (which we

[11] An identification can be said to be unfinished, in the sense we mean, if it is nonconstructive or if the identification of its properties, or its property's properties, etc., is nonconstructive. A restriction that prohibits nonconstructive identifications thereby prohibits unfinished ones.

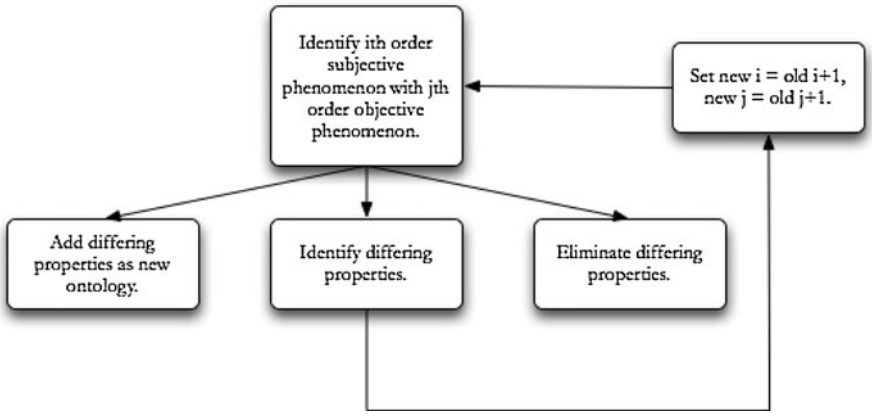

Fig. 2 The essential structure of the problem. M is forbidden from adding and eliminating ontology, so must loop back to a structurally identical problem at the immediately higher order

will call 'greenishness'). For the afterimage seems to be green which, given PI, is sufficient for the subjective correspondent of the afterimage to be greenish.

M, however, does not already regard the *circuit state* C as being greenish prior to the time t of the identification. That's because *greenishness*, given PI, is a property that applies only to subjective phenomena (and M is always diligent about such things). So M can only have already applied *greenishness* to C if C has already been identified with something subjective.[12] But since this is the first time that *any* objective phenomenon has been identified with something subjective, C cannot have already been identified with anything subjective. Hence C is not already regarded as *greenish.* Hence, at the time M comes to make the identification, M regards the subjective correspondent of the afterimage as having a property that he has not ascribed to C. This must be addressed at the time the identification is made.

There are three ways to address it:

M must decide that the correspondent of the afterimage is not really greenish after all or
M must decide that the circuit-state C is greenish after all, where greenishness is a property that is *not* discoverable by third person methods, or
M must decide that C is greenish, where greenishness *is* discoverable by third person methods. I.e., M must identify the greenishness with some third-person discoverable property.

The first option, however, would violate M's first restriction, since it forces M to admit he was wrong about a subjective correspondent. And the second option violates the second restriction, since it forces M to regard the greenishness as something outside of objectively discoverable reality. So, M must opt for the third, and identify greenishness with some objective property of the circuit state. However, therein is the contradiction, since the identification of the subjective

[12] Nor, for the same reason, could M have already identified greenishness with some property that already applies to C.

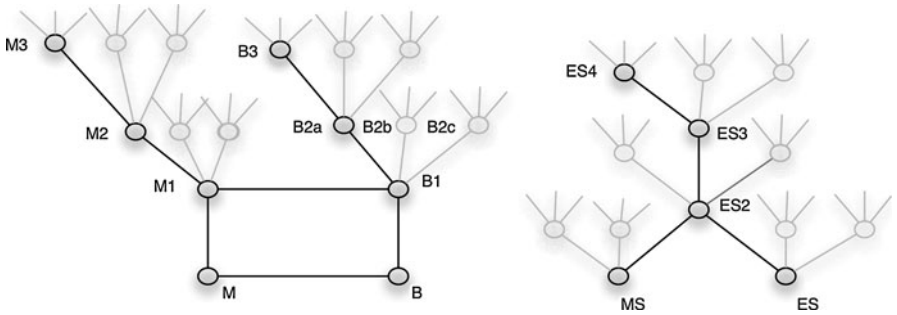

Fig. 3 The *left hand diagram* shows the problem M faces. The *horizontal bars* represent the reductio assumption that M has identified M (mind) and B (body), and, for good measure, their properties M1 and B1. Though it deductively follows that each property of M1 is a property of B1, the specific identities do not deductively follow since (e.g.,) M2 could be identical to any of B2a, B2b or B2c. And even if M2 is identified with B2a, the problem re-arises at the next order, and so on to infinity. Wherever M stops, he must put precise *bars* between two nodes without putting *bars* between their properties. On the *right*, why M does not face the problem when identifying the morning and evening star. Since M is able to *add* the property ES2 to MS (i.e., by adding a leg between MS and ES2) he effectively *transplants* the tree above it to become part of MS's property tree. There is therefore no need to identify higher order properties like ES4 with any higher order property of MS, since ES4 was transplanted along with ES2

correspondent with the circuit state is supposed to be the highest ordered subjective–objective identification completed thus far, yet, in order to make it, M must complete a prior or simultaneous higher-order identification (Fig. 2).

The Demonstration

We just need to show that these considerations are appropriately general. We start as we did before, assuming for reductio that M succeeds in making some identification. It follows that some SX^i was both the (joint) first and the (joint) highest ordered subjective phenomenon to be identified with some objective phenomenon Y^j. In order that we can use the present tense let us imagine that we are speaking at the time that M makes this identification and that the subjective phenomenon SX^i is tokened at the time of speaking.

Since M tokens SX^i, where SX^i is that state (according to PI) that M enters only when X^i seems some way or other, it follows that X^i seems some way or other; and if X^i seems some way or other then it must seem to have a property X^{i+1}.[13] It seeming to

[13] For X^i to seem some way or other is to seem discernible from the other ways it might have seemed, a truth that holds even if it is a property. But X^i can only seem discernible from other ways it could have seemed if it itself seems to have properties. X seems discernible from Y, after all, only if the properties of X themselves seem discernible from those of Y.

A slightly different argument to the same conclusion: Either X^i is a property or not. Suppose not. In that case, since it seems some way or other, it must seem to have some color, be some size and shape, and so on. Hence X^i seems to have properties. If, on the other hand, X^i is some seeming property of X^{i-1} then X^i itself must seem to have the property of *modifying* X^{i-1}; of *being apparent;* of *being a property*; of *marking* X^{i-1} *as distinct from things that lack* X^i; etc. It will, moreover, have idiosyncratic properties too. E.g., if X^i is the property (say) of *being green*, it will seem to have the property of *marking its bearer as colored*; of *making its bearer visually distinguishable from red things*; of *camouflaging its bearer against green backdrops*; and so on. So again, if X^i seems some way, then it seems to have properties.

have a property X^{i+1} implies, in turn, that SX^i has SX^{i+1} by PI.[14] But the always diligent M does not already believe that Y^j has SX^{i+1} since SX^{i+1}, according to the principle PI that introduced it, is only a property of *subjective* phenomena and Y^j has, until now, been regarded as a purely *objective* phenomenon (this being M's (joint) *first* subjective–objective identification). Hence M, in order to identify SX^i with Y^j, must also come to believe either that Y^j has SX^{i+1} or that SX^i does not have it after all. But he cannot deny that SX^i has SX^{i+1} without surrendering his claim to subjective infallibility. And he cannot add SX^{i+1} to Y^j as an *additional* property *over and above* the properties Y^j can objectively be discovered to have, for that violates M's commitment to everything being objectively discoverable. Hence he must *identify* SX^{i+1} with some objectively discoverable property Y^{j+1} of Y^j. But if he does that, then we reach contradiction, since by assumption SX^i and Y^j were supposed to be the highest order subjective/objective phenomena to be successfully identified.

Since the assumption that he makes some identification leads to contradiction, it follows that M cannot identify any subjective phenomenon with any of the states, objects, events, properties or processes, either abstract or concrete, that appear in his objective picture of the world.

Corollary M faces the same problem when he tries to identify *himself* (i.e., the experiencer of the subjective states) with something objective. To see why, note that it's not obvious to M that he is the whole of the hardware he sees in the mirror, since there are other things he could possibly be. For example, he might be a special part of the hardware—perhaps the motherboard, or the CPU. Or he might be a process run by the hardware; a running application or virtual machine. And to identify himself with, say, the motherboard consistent with Leibniz's Law M must come to believe that the motherboard enters all the states that he does, hence that either the motherboard enters state SX^0 or that, in fact, M does not. But, given his first restriction, M will not remove SX^0 from the states that M himself enters into, as if it never seemed to him that X^0; and, given his second, he will not add SX^0 as a non-objective state of the motherboard. So, if he is to avoid nonconstructive identifications, that leaves only the possibility of identifying SX^0 with some *specific* objective state of the motherboard. But this task we know M cannot complete, from the proof above.

Relaxing the Soft Restrictions

Obviously the robot can escape the problem if he can violate the restrictions that originally bound him to it. What is interesting, here, is that each possible violation corresponds to a human approach to the mind–body problem, suggesting that what we have here is a problem in the same class as the problem of consciousness, but discovered in an artificial context.

[14] Moreover, M infers that it does. For M is positioned to know that X^i seems to M to have X^{i+1}—or else it does not really seem that way to M after all—and M can infer from this, together with his rule PI, that SX^i must have SX^{i+1}. Since we are assuming that M infers the properties of potential identificanda whenever he is in a position to, it follows that he does.

For example, if the first restriction is relaxed then M can engage in eliminativist strategies. Just as the qualia eliminativist[15] denies that the experience really has the qualitative, phenomenal properties that intuitively make the experience difficult to reduce, M could deny that the subjective phenomenon SX^i really has the troublesome property SX^{i+1}. However, for reasons similar to those that oppose the eliminativist, M may find this position implausible. For, given PI, denying that SX^i has SX^{i+1} amounts to denying that X^i even *seems* to have X^{i+1}. Hence M must either wriggle free of PI, which was formulated as a definitional truth, or deny the obvious way things seem.

Relaxing the second restriction, meanwhile, permits M to add the property SX^{i+1} that Y^j seems to lack as a *new* property over and above those discoverable by objective third-person investigation. This again would remove it as an impediment to the SX^i/Y^j identification, but at the cost of implying something like property dualism, wherein subjective properties (like qualia) are simply added to objective phenomena (like brains) that seem to lack them, while being considered ontically distinct from any third-person observable properties.[16] Alternatively, M might violate the second restriction, not by adding properties so as to facilitate identification, but by simply eschewing identification altogether and counting the correspondent of the afterimage, SX^0, with all its properties, as a brand new element in his ontology. This corresponds to the traditional substance dualism associated with Descartes.[17]

By relaxing the third restriction M gets to leave some identification unfinished; where we count M's identification as unfinished if he identifies X and Y while leaving questions about their apparently different properties unanswered. Unfinished identifications are common human responses to our own problem of consciousness. We get an unfinished identification, for instance, if some researcher identifies pain with lamina I neurons stimulating parieto-insular cortex, but doesn't identify the pain's unpleasantness with any property of this brain event. Equally, if the researcher plows ahead and identifies the unpleasantness with (say) the lamina I stimulations' tendency to spread activity into the anterior cingulate, she will also have to say, if she is to make sense, that though the unpleasantness *seems* discernible from the effect on anterior cingulate—i.e., though it seems to have different properties—each property of the one is in fact identical to *some* property of the other. But supposing she stops there, without providing any further specifics, this is again an unfinished identification.[18,19]

[15] See e.g., Dennett (1988), Rey (2007).

[16] See Jackson (1982) and Chalmers (1996) for recent property dualist positions.

[17] To get something closer still to Cartesian dualism, M could also pursue the same strategy in response to the corollary, holding that he is not the hardware that he sees in the mirror, nor any other objective thing he finds in the physical world.

[18] It might be thought that even when relaxing the third restriction M might prove, using just the first two restrictions, that no specific identification is possible in principle, hence hankering after a future finished theory would make no sense. This tempting thought is only half right. M could make specific but *unfinished* identifications by identifying A with B while only committing to each of A's properties being identical to *some property or other* of B. In this way a non-specific identification at order O + 1 permits an unfinished but specific identification at O.

[19] This tendency to halt and leave the matter unfinished, typically at a very low order, may plausibly explain why the human problem of consciousness does not obviously present itself as a regress.

Objections and Responses

Objection 1 Perhaps M could justify the identification in a different way, without using Leibniz's Law.

Response It is a mistake to think that M is trying to *justify* the identification *using* Leibniz's Law. Leibniz's Law, rather, is what is getting in the way. For satisfying Leibniz's Law is a *necessary* condition for identifying phenomena, not a *sufficient* one. And since it is a necessary condition, it must be satisfied no matter what identification strategy M uses.

A surprising consequence of this is that even an *a priori* deductive argument for the identity of some subjective with some objective phenomenon may not be enough to get M past his problem. For all that would succeed in doing is giving M the strongest possible argument that $SX^i = Y^j$ even though M holds SX^i to have properties that he does not hold Y^j to have, and even though he cannot find a tractable way of changing his mind. Imagine being in the same situation; with an apparently deductive argument that Smith = Jones, but no way to rationally relinquish your belief that Jones is male and Smith female. You would encounter 'how possibly' questions—*how could Smith possibly be Jones when they are of different genders?*[20]—and you may well begin to doubt the deductive argument.[21] M's problem, in the same way, arises no matter the strength of the argument for identification. A deductive argument for identification would merely lead to paradox and puzzlement, not resolution.

Objection 2 Doesn't this problem generalize badly? If it were true that M had to justify the identification of every property of every property, to infinity, then wouldn't M have to do that for *all* identifications, not just those that involve the identification of the subjective and the objective? Since humans have no problems identifying such things as (say) the morning star and the evening star, doesn't this make the problem disanalogous to the one humans face?

Response It would, but in fact M can identify such things as easily as humans do, without violating his restrictions. The thing to note, in understanding why, is that M is not forced to identify the properties of the properties of…. each identificandum in order to make any arbitrary identification. He is only forced to do so when (a) the properties are not already held to agree and (b) the options of adding or subtracting the differing properties are always unavailable. But when M comes to identify *being an unmarried man* with *being a bachelor*, he may already hold the two to have all the same properties; e.g., to both apply solely to males, to both apply solely to humans,

[20] Compare, obviously, with the question: 'How could pain possibly be nothing more than c-fibers firing?' Identifications which proceed while accompanied by such mystification are dubbed 'gappy' identifications by Levine (2001), who regards them as the typical reaction to phenomenal-physical identifications. Such 'how possibly' reactions, I suggest, arise from the fact that the identification hasn't been properly squared with Leibniz's Law.

[21] Paradoxes illustrate that (apparently) deductive arguments only rationally compel belief when the conclusion is appropriately acceptable. Zeno's paradox, for example, was an apparently sound deductive argument to the effect that bodies could not move. Those who historically struggled to deny either the premises or deductive validity of the argument were nonetheless rational to resist its conclusion.

etc. In this case M can identify without any issue. Alternatively, to identify the morning star and the evening star, or water with H_20, where they don't already seem to have the same properties, M merely needs to add *appears in the evening* to the morning star, *appears in the morning* to the evening star, *has a micro-rigid structure* to water, etc. Nothing stops M from adding *has a micro-rigid structure* to his concept of water, as an additional property over and above its *commonsense* properties, the way the second restriction stops M from adding *is greenish* to his concept of the motherboard, as an additional property over and above its *objective* properties.[22]

Thus M's problem, like our own hard problem of consciousness, arises only with respect to his subjective representational states. It does not arise for ordinary objective phenomena.

Objection 3 Doesn't the derivation tacitly depend upon the following claim?

> *Endless Justification (EJ)*: For P to be justified, all of the logical consequences of P must be antecedently (or independently) justified. E.g., For the proposition 'the sea is blue' to be justified, the proposition 'the sea is colored' must be justified first.

And isn't EJ an excessive demand, for it ensures that no proposition could ever be justified? Isn't it much more plausible, in contrast, that *by* justifying P, one *thereby* justifies its deductive consequences?

Response It is true that if M's problem depended on EJ, then he would never be able to justify any proposition, and hence never justifiably identify anything with anything. Yet in response to the last objection we saw that, consistent with all his restrictions, M *can* justifiably identify all kinds of things. So it seems that M's problem cannot depend on EJ after all.

One way to see that M's problem does not rely on EJ is to grant that

1. M need not antecedently or independently justify the deductive consequences of each identification he makes. By justifying the identification, he *thereby* justifies all its consequences.

thereby falsifying EJ by decree, while also granting that

2. M succeeds in identifying some subjective and some objective phenomenon.

If the contradiction is still derivable, then the derivation obviously did not make use of EJ.

The contradiction, indeed, still follows, pretty much the same as before. Given (2) there must be some time t at which M made his first identification, or batch of identifications, and some subjective phenomenon SX^i must have been the (joint)

[22] Subtraction is another option. To identify heat with molecular motion, for example, humans may *either* add the 'feeliness' of the heat as a new property of the molecular motion *or* just as plausibly remove the 'feeliness' from the heat by re-categorizing it as a property of the human's own sensory system, not of the heat itself. Since M claims no infallibility when it comes to objective phenomena like heat, he can, consistent with his first restriction, revise away those of its properties that make it difficult to reduce.

highest ordered identification in that batch *that actually involved some work* (this is the only real amendment to the original argument) with all the ones at higher orders, if there are any, being cognitive freebies of the sort M gets from (1). Since M always identifies consistent with Leibniz's Law, it follows that:

3. M accepted at t that each property of SX^i was identical to *some* property of Y^j.

But we know from the third restriction that M does not ever identify non-constructively. Hence

4. If M accepted at t that each property of SX^i was identical to *some* property of Y^j, then, at t, M held each *specific* property of SX^i to be identical with some *specific* property of Y^j.

The consequent of (4) follows. But in that case how did M identify each *specific* property of SX^i with some *specific* property of Y^j? These *specific* identifications are not cognitive freebies that come along without effort as deductive consequences of the identification of SX^i with Y^j. It does not follow, for example, that if one of them has properties P, Q and R and the other has F, G and H, that P = F, since P might just as well be G or H. So to make the *specific* identifications implied by the consequent of (4), M must have done *additional work*. But this contradicts the assumption that the identification of SX^i with Y^j was the (joint) highest ordered identification that actually involved any.

It helps, here, to compare M's problem with the human analog. Imagine that a researcher identifies pain with lamina I neurons stimulating parieto-insular cortex. As a deductively entailed freebie, she gets to say that each property of the one is identical to some property of the other. But she does *not* get to identify the pain's unpleasantness with (say) the lamina I stimulations' tendency to spread activity into the anterior cingulate. That *specific* identity does not deductively follow, and so requires extra work. Hence she must either do the work or leave the reductive project unfinished.[23]

Objection 4 The subjective correspondent of the fire truck was supervenient upon M's functional inability to discriminate illusory from veridical states. So isn't it *obvious* that M's subjective states are functional? Why, then, can't M make this trivial observation? This needs to be explained.

Response To make the observation, M would have to reason the way the objector just did. He would look at himself from the third person perspective (we can imagine him using a mirror or a blueprint) and see a piece of hardware. He would then note that it is *obvious* that the subjective correspondents of the hardware, if it

[23] Contrast this with how M (or a human) is able to identify (say) the morning star with the evening star. Assuming they only ever differed with respect to one property, after adding the property *appears in the morning* to the evening star and the property *appears in the evening* to the morning star the identification can go ahead. Why is there no regress? Because all the *specific* properties of properties of… the morning star and properties of properties of… the evening star end up being the same, to infinity, when the differing properties are *added*. To see why, let the evening star's 'property tree' be the structure of its properties, its properties' properties etc., branching up to infinity. Now take an arbitrary property P* that appears in this property tree somewhere in the sub-tree above the property P that was added to the morning star. Since P was *added,* the whole of the sub-tree to which it leads is added too, and so P* appears in the morning star's property tree after all. (see Fig 3.).

has any, are mere functional states, supervening on the hardware's inability to make certain discriminations. Finally, all M needs to do is decide that *he* is the hardware he sees in the mirror, and he is in a position to make the inference. Indeed, if he gets to know the hardware well enough, it seems that he may even be able to identify each of his subjective correspondents with some specific state of the hardware, using an argument like the following:

1. SX^1 is that unique property I use to mark my subjective correspondents as correspondents of (apparent) red things.
2. Y^1 is that unique property the hardware uses to mark its subjective correspondents as correspondents of (apparent) red things.
3. I am identical with the hardware.
4. Therefore SX^1 is Y^1.

The first problem with this, however, is that it is just a valid deductive argument for the conclusion that SX^1 is Y^1—but we have already argued, in response to objection 1, that such an argument can be of no help to M. The fact that he has a great argument that SX^1 is Y^1 does not change the fact that M attributes different properties to SX^1 and Y^1 and can find no finite way to change his mind. The best M can get from a compelling argument is a 'how possibly' paradox wherein he has great evidence that $SX^1 = Y^1$ yet no way to see how the identity is possible.

The second problem is that M could not accept the premises of this argument in any case, since our corollary tells us that M cannot accept premise (3). No similar argument is likely to fare any better, moreover, since they would all require some premise just like it. M, after all, has little reason to identify his state with some state of the motherboard (or the CPU, or the virtual machine, etc.) unless he has some reason to identify *himself* with the motherboard (or the CPU, etc.).[24]

Objection 5 We can't be sure that the problem is unsolvable. Rather than plowing through an intractable space, identifying the properties of the properties and so on, M might discover an ingenious proof that grants him all the specific identifications to infinity at a stroke. Or he might find that the search space is in fact finite, with the same properties appearing at the higher orders that already appeared at the lower.

Response The project is not to show that the problem is unsolvable in any way, but to show that the identification of subjective and objective phenomena cannot be completed in the regular, procedural way that we identify ordinary things like the morning star and the evening star. It is therefore difficult in a way that identifying other things is not.[25]

[24] Indeed, without a premise like (3), M has not even got the weaker argument to the effect that his subjective correspondents are identical to *some* functional state or other of the motherboard, or CPU, etc.

[25] I thank Matthias Scheutz for pushing the possibility of an ingenious proof, as well as the possibility of a looping property space. Note, though, that no proof of a single identity is sufficient to escape the problem—see objection 1. What's needed is a proof of infinitely many specific identities, and the mind boggles at what that might look like. We should also be pessimistic about the possibility of a repeating property space, since the property *is a property of a property of a property of a red thing* cannot logically appear at any level lower than the 3rd; and there are many properties of this sort at every order (including some more interesting ones like *is the horribleness of the hunger-feeliness of the biological state of a badger*).

Objection 6 But why couldn't M's creator solve M's problem, by *telling* him that he is the motherboard, or the CPU, or by otherwise pre-programming the identifications that M struggles to make? Couldn't she simply enter the relevant identities into the system, creating an M without any problems?

Response No. For suppose that she did. Then we could simply take the highest order identity that the programmer coded into M and prove that there must have been a higher one. The contradiction would be derivable just the same. That's not to say that the programmer cannot program the identity into the machine, as if, upon trying, she would be hit by a thunderbolt. Rather, the claim is that the programmer cannot program the identity into the machine *while keeping M in good standing with Leibniz's Law*. If the programmer were to try to code the relevant identifications into M while observing M's restrictions, she would have to stop somewhere and, at whichever order she stops, she violates Leibniz's Law on M's behalf. For she would have given M a brute belief in the identity of some subjective phenomena whose properties M does not yet hold identical.[26,27]

But in any case, our aim is not to create a problem-free robot, as if this were an engineering project and we simply needed to get the thing working. Rather, we want to better understand our own hard problem through contemplating M's. And we do this best by not getting involved. For even if we could intervene to free M from his problem, that wouldn't change the fact that when we *don't* intervene, M is stuck. And that gives us insight into our own predicament, where we have no programmer to solve our problems for us.

Concluding Remarks

By attempting to construct an intelligent system that contemplates its own illusions, we create a problem that resembles our own hard problem of consciousness. The problem, roughly, was that the robot could not make a subjective–objective identification without making a different one first. The robot could avoid the regress only by supposing the recalcitrant features to be new, ontically additional, aspects of reality, or by removing them from its ontology entirely, or by putting its physicalist theory into abeyance. But all of these responses resemble classic approaches to our human problem of consciousness.[28]

[26] A programmer who knew the specific identities to infinity might enter a rule that generates them for M. Even given such a wonderful rule, though, the identification still appears intractable. For M must still identify the properties PX and PY of X and Y (using the rule) before identifying X and Y, and the properties PPX and PPY of PX and PY (using the rule again) before identifying those, and so on. There is no starting point to such a process even if each identification takes only one step.

[27] The programmer could just give M the same representation for both identificanda thus saving it from even having to identify them. However, in that case the robot would not be able to contemplate them being distinct, since it needs separate representations to do that. We would have yet another machine that dodges its hard problem by being insufficiently smart.

[28] Indeed, it is perhaps because we each take on of these strategies for avoiding the regress and try to make it work (e.g., by taking the dualist option and then struggling with mental causation) that explains why we get the problems associated with the ways out of the regress, not the problems associated with the regress itself; hence why it is that the problem of consciousness does not seem to present itself as a regress.

M's problem, though, was purely algorithmic—a procedural problem in which each identification task always presupposed another. To the extent that our own hard problem resembles it, we have new reason to resist responding with extravagant metaphysical proposals. Yes, we can identify A with some seemingly different B by being a dualist or eliminativist about the seeming difference and, no, we can't make the identification if we forbid ourselves such strategies; but that might only be because we can't *finitely complete* the identification task, not because our stance is *incorrect*. M himself, after all, was presumably *right* to *try* to reduce himself to a purely physical, entirely ordinary objective item, but that didn't mean he could do it. It may be that we face a similar problem.

Acknowledgments My thanks for helpful comments go to Bert Baumgaertner, Joel Friedman, Jimmy Licon, Paul Teller, and to participants at the 11th annual meeting of the International Association for Computing and Philosophy, with particular thanks to Matthias Scheutz; and finally to several anonymous responders at *Minds and Machines* and my tenure review.

References

Chalmers, P. (1996). *The conscious mind.* Oxford: Oxford University Press.

Dennett, D. (1988). Quining qualia. In A. Marcel & E. Bisiach (Eds.), *Consciousness in modern science*. Oxford: Oxford University Press.

Jackson, F. (1982). Epiphenomenal qualia. *Philosophical Quarterly, 32*, 127–136.

Levine, J. (2001). *Purple haze: The puzzle of consciousness*. Oxford: Oxford University Press.

Minsky, M. (1965). Matter, mind and models. *Proceedings of International Federation of Information Processing Congress, 1*, 45–49.

Nagel, T. (1974). What is it like to be a bat? *Philosophical Review, 83*(4), 435–450.

Rey, G. (2007). Phenomenal content and the richness and determinacy of colour experience. *Journal of Consciousness Studies, 14*(9–10), 112–131.