**Daniel Dennett - What RoboMary Knows**[[1]](http://ase.tufts.edu/cogstud/dennett/papers/RoboMaryfinal.htm" \l "_ftn1" \o ")

            Frank Jackson’s thought experiment about Mary the color scientist is a prime example of an intuition pump, a thought experiment that is not so much a formal argument as a little scenario or vignette that has been pumping philosophical intuitions with remarkable vigor since it first appeared in 1982. For sheer volume and reliability, this must count as one of the most successful intuition pumps ever devised by analytical philosophers.  But is it a good intuition pump?  How could we tell?  Douglas Hofstadter’s (1981) classic advice to philosophers confronted by a thought experiment is to treat it the way scientists treat a phenomenon of interest: vary it, turn it over, examine it from all angles, and in different settings and conditions, just to make sure you aren’t taken in by illusions of causation. During the last twenty years, philosophers have examined many variations and defended many different responses, but they have singularly neglected some of the possible settings of the knobs.  More than a decade ago, I conducted a preliminary exploration of the knobs, and issued a kill-joy verdict that has been almost universally disregarded: “Like a good thought experiment, its point is immediately evident even to the uninitiated. In fact it is a bad thought experiment, an intuition pump that actually encourages us to misunderstand its premises!” (Dennett, 1991, p398) In fact it is much more difficult to imagine the scenerio correctly than people suppose, so they imagine something easier, and draw their conclusions from that mistaken base. In an attempt to bring out the flaws in the thought experiment, I encouraged people to consider a variant ending:

And so, one day, Mary’s captors decided it was time for her to see colors. As a trick, they prepared a bright blue banana to present as her first color experience ever. Mary took one look at it and said ‘Hey! You tried to trick me! Bananas are yellow, but this one is blue!” Her captors were dumfounded. How did she do it? “Simple,” she replied. “You have to remember that I know *everything–*absolutely everything–that could ever be known about the physical causes and effects of color vision. So of course before you brought the banana in, I had already written down, in exquisite detail, exactly what physical impression a yellow object or a blue object (or a green object, etc.) would make on my nervous system. So I already knew exactly what *thoughts* I would have (because, after all, the ‘mere disposition’ to think about this or that is not one of your famous qualia, is it?). I was not in the slightest surprised by my experience of blue (what surprised me was that you would try such a second-rate trick on me). I realize it is *hard for you to imagine* that I could know so much about my reactive dispositions that the way blue affected me came as no surprise. Of course it’s hard for you to imagine. It’s hard for anyone to imagine the consequences of someone knowing absolutely everything physical about anything!’” (p399-400)

            It is standardly assumed that things could not proceed this way. As Jackson disarmingly put it in the original article, “It seems just obvious that she will learn something about the world and our visual experience of it.” (p128)  That, I claimed, is a mistake, and that is what is wrong with Mary as a thought experiment. It just feels so good to conclude that Mary has a revelation of *some* sort when first she sees color that nobody wants to bother showing that this is how the story must go. In fact it needn’t go that way at all.  My variant was intended to bring out the fact that, absent any persuasive argument that this could not be how Mary would respond, my telling of the tale had the same status as Jackson’s: two little fantasies pulling in opposite directions, neither with any demonstrated authority.  I thought that I had said enough to make my point, but a decade of further writing on Mary by many philosophers and their students have shown me that I should have been more patient, more explicit, in my objections. I underestimated the allure of this intuition pump by a wide margin. So I am returning to the fray, and this time I will make my case at a more deliberate pace, dotting the i’s and crossing the t’s.

            First, I have found that some readers–maybe most–just didn’t get my blue banana alternative.[[2]](http://ase.tufts.edu/cogstud/dennett/papers/RoboMaryfinal.htm" \l "_ftn2" \o ") What was I saying? I was saying that Mary had figured out, using her vast knowledge of color science, *exactly what it would be like for her to see* something red, something yellow, something blue in advance of having those experiences.[[3]](http://ase.tufts.edu/cogstud/dennett/papers/RoboMaryfinal.htm" \l "_ftn3" \o ") I asserted this flat out—*in your face,* as it were—in order to expose to view the fact that people normally assume that this is impossible on the basis of no evidence or theory or argument, but just on the basis of ancient philosophical tradition going back at least to John Locke. Perhaps a little dialogue would help bring out the intended point:

TRAD: What on earth do you mean? *How*could Mary do that?

DCD: It wasn’t easy. She deduced it, actually, in a 4765-step proof (for red–once she’d deduced what red would look like to her, green fell into line with a 300-step lemma, and the other colors–and all the hues thereof–were relatively trivial extensions of those proofs).

TRAD: You’re just making all that up! There are no such proofs!

DCD: This is a thought experiment; I get to make up all sorts of things. Can you prove that there are no such proofs?  What established fact or principle am I contradicting when I help myself to a scenario in which she deduces what colors would look like to her from everything she knows about color?

TRAD: Look. It’s just obvious! *You can’t deduce what a color looks like if you’ve never seen one!*

DCD: That’s an interesting folk theorem, I must say. Here’s another: If you burp, sneeze and fart all at the same time, you die. Both sound sort of plausible to me, but is there any scientific backing for either one of them?

1. She’ll be surprised, dammit!

            If the Mary thought experiment was intended simply to draw out and illustrate vividly the implications of a fairly standard way of thinking that many, probably most, people have, it might be a useful anthropological exercise, an investigation of folk psychology laid bare. But those who have championed Mary have thought that it might actually prove something bigger, not just the conclusion that most people’s unexamined assumptions imply dualism—I think we already knew that, but maybe not—but the conclusion that dualism is true!  The fact that philosophers would so much as *entertain* such an interpretation of such a casual exercise of the imagination fills me with astonishment. I had no idea philosophers still put so much faith in the authority of their homegrown intuitions. It is almost as if one thought one could prove that the Copernican theory was false by noting that it “seems just obvious” that the earth doesn’t move and the sun does.

            Consider, for instance, the recent article  “Mary Mary Quite Contrary,” by George Graham and Terence Horgan (2000) Graham and Horgan (G&H henceforth) have usefully managed to distill precisely the unargued intuition that I have been attempting to isolate and discredit for fifteen years and more–the one we might express as “She’ll be surprised, dammit!” G&H begin by distinguishing two main materialist responses to Mary: thin and thick materialism. Thin materialism, of which I am one of the few exponents, denies that Mary learns anything post-release. Thick materialists attempt to salvage materialism while going along with the gag that Mary is startled, delighted, surprised, or something like that, when she is released from her colorless captivity. G&H’s strategy is first to declare briskly that thin materialism is a non-starter in need of no refutation since it “has been amply criticized by others,” (p63). The only critics they list are McConnell (1994) and Lycan (1996). Since I replied at some length to McConnell in the same journal (Dennett, 1994), and since Lycan doesn’t criticize my version of thin materialism, I don’t find this criticism ample, but I must admit that G&H are only going along with the mainstream in ignoring my brand of thin materialism. That’s why the current essay is necessary.

            G&H spell out the best of the thick materialist campaigns–Michael Tye’s PANIC–and imagine their own variation on the original theme: Mary Mary, the daughter of the original Mary, and a devotee of Tye’s brand of thick materialism. According to Tye’s PANIC theory, “phenomenal character is one and the same as Poised Abstract Nonconceptual Intentional Content” (Tye, 1995, p137), which means roughly that it is content that is “in position to make a direct impact on the belief/desire system” and is about non-concrete, non-conceptualized discriminable properties.  It follows from Tye’s view, they claim, that Mary Mary, upon release, *shouldn’t be surprised.* As they say, “In the end, Tye’s version of thick materialism is just *too thin*. And this problem threatens to arise for any materialist treatment of phenomenal content.” (p77).

            I had previously viewed Tye’s alternative to my brand of thin materialism as giving too much ground to the qualophiles, the lovers of phenomenal content, but thanks now to G&H I can welcome him into my underpopulated fold as a thin materialist *malgré lui*, someone who has articulated much more painstakingly than I had just what sorts of functionalistically explicable complexities go to *constitute* the what-it-is-likeness, the so-called *phenomenality*, of conscious experience. I applaud G&H’s analysis of Mary Mary’s predicament,  leading inexorably to the conclusion that since she already knows all the facts, has all the information needed to have anticipated *all*the noticeable, remarkable-upon properties of her debut experience in a colored world, she should not, in spite of what Tye claims, be (or expect to be) surprised. Here, in a nutshell, is what they say:

First, what is psychologically significant about the PANIC properties is just the functional/representational role they play in human cognitive economy–something that Mary thoroughly understands already, by virtue of her scientific omniscience. . . . . Second, what is psychologically significant about phenomenal concepts (given Tye’s theory) is that they are *capacity-based* concepts; . . . . But she already understands these capacities thoroughly, including how PANIC states subserve them, even though she does not possess the capacities herself.  No expected surprises there, either.

            Third, the psychological distinctiveness of beliefs and knowledge-states employing phenomenal concepts is completely parasitic (given Tye’s theory) upon the capacity-based nature of the phenomenal concepts.  So she already understands well the *nature* of these beliefs and knowledge-states . . . . So Mary Mary, as a True Believer in Tye’s PANIC theory of phenomenal consciousness, has no good reason to expect surprise or unanticipated delight upon being released from her monochrome situation. (p71-2)

            In short, Tye should join me in predicting that Mary Mary, like her mother Mary, would *not* be surprised or delighted at all. She’s been there, done that, in her vast imagination already, and has nothing left to learn.  So what’s the problem? Why don’t G&H join Tye and me?  (I’m presuming for the fun of it that Tye is now on my side.) Because–and here comes the super-pure, double-distilled intuition that I’ve been gunning for–“Surely, we submit, she should be both surprised and delighted.” (p72)  “Surely.”  As I noted in “Get Real” (Dennett, 1994) in one of my many commentaries on Ned Block, “Wherever Block says ‘Surely,’ look for what we might call a mental block.” (p549). Block is perhaps one of the most profligate abusers of the “surely” operator among philosophers, but others routinely rely on it, and every time they do a little alarm bell should ring. Here is where the unintended sleight-of-hand happens, whisking the false premise by the censors with a nudge and a wink. G&H do pause momentarily to ask why they are so sure, and this is what they answer:

What will surprise and delight Mary Mary . . . . is (it seems to us) the unanticipated *experiential basis* of her concept-wielding, recognitional/discriminatory, capacities and the acknowledged richness of her experience; she never expected polychromatic experience to be like *this*. (p72)

I know that it seems to many people that there is this extra “richness,” this “*experiential basis*” over and above all the PANIC details, but I have claimed that they are just wrong about this, and I have offered a diagnosis of the sources of this deep-seated theorists’ illusion.  In “Quining Qualia” (Dennett, 1988), I discussed  the example of the torn Jell-O box, half of which has shape property M, and the other half of which is the only *practical* M-detector: the shape may *defy description*but it is not literally ineffable or unanalyzable; it is just extremely rich in information. It is a mistake to inflate practical indescribability into something metaphysically more portentous, and I have been urging people to abandon this brute hunch, tempting though it may be. G&H cannot bring themselves to abandon the intuition, butmore important, they cannot even bring themselves to acknowledge that their whole case thus comes down to simply announcing their continued allegiance to a claim that, whether it is true or false, has been declared false and hence could use some support. They offer no support for it, but they do keep coming back to it, again and again:

            Although phenomenal states may indeed play a PANIC role in human psychological economy, their phenomenal character is not reducible to that role. It is something more, something surprising and delightful. (p73)

Who says?  This is just what I have denied, at length.

            Its greater richness is what is surprising and delightful about it, and Tye’s theory leaves this out. (p73)

This “greater richness” is just what needs to be demonstrated, not assumed. After all, the point of the Mary example is supposed to be that although thanks to her perfect knowledge she can anticipate *much* of what it will be like to see colors, she cannot anticipate it *all*. Since some of us have claimed that there is no reason to deny that all the “greater richness” *is* accessible to Mary in advance, this bald assertion by G&H is question-begging. It simply won’t do to lean on the obvious fact that under normal circumstances, indeed under any circumstances except the wildly improbable extreme circumstances of this thought experiment, Mary would learn something.

But she *will*experience surprise and unanticipated delight, upon release from her monochromatic environment–which presumably should lead her to repudiate the materialist theory she previously accepted. (p74)

            So they say. Now thin materialism may, in the end, be false, but you can’t argue against it by just saying “Surely not!”  I have claimed that the richness we appreciate, the richness that we rely on to anchor our acts of inner ostension and recognition is *composed of* and *explained by* the complex set of dispositional properties that Tye has called PANIC properties. G&H make the mistake of assuming that there is, in addition to all this, a layer of “direct acquaintance” with “phenomenal properties.” They say baldly:

There is also direct acquaintance with phenomenal character itself, acquaintance that provides the experiential basis for those recognitional/discriminatory capacities.(p73)

And also:

She claims to be delighted . . . . Auto-phenomenology suggests strongly, *very*strongly, that she is right about this: the intrinsic phenomenal character of color experience is distinct from, and provides the basis for, these recognitional/discriminatory capacities. (p77)

This is, according to me, just about backwards. These capacities are themselves the basis for the (illusory) belief that one’s experience has “intrinsic phenomenal character,” and we first-persons have no privileged access at all into the workings of these capacities. That, by the way, is why we shouldn’t do auto-phenomenology. It leads us into temptation: the temptation to take our own first-person convictions not as data but as the undeniable truth.

So on his [Tye’s] story, Mary Mary’s post-release heterophenomenological claims evidently must be viewed as rationally inappropriate, and thus as embodying some kind of error or illusion. *That*is the basic problem: the apparent failure to provide adequate theoretical accommodation for the manifest phenomenological facts. (p77)

The basic problem, they say, is dealing with these “manifest” facts, but it’s only a problem if, in fact, she will learn something. It is not a problem for my view (and Tye’s, if he’ll join the thin materialists); she won’t learn anything, and won’t be surprised; there are no such manifest phenomenological facts. At this point, if you are like many of my students, you are beset with frank incredulity. *Of course* Mary learns something on release! She *has to!* Oh? Then please give me an argument, based on premises we can all accept, that demonstrates this. But I have never seen such an argument even attempted. “It stands to reason!” people say, and then they decline to offer any reasons, thinking them somehow uncalled for. I call for them.

            In response to the previous paragraph in an earlier draft, Bill Lycan has answered the call:

Here's a way to see why some of us think Mary does learn something.  What one knows when one knows w.i.l. [what it's like] to experience a blue sensation is ineffable; at least, it's very tough to put into (noncomparative) words. One resorts to the frustrated demonstrative: "It's like...*this*."  The reason physically omniscient Mary doesn't know what it’s like is that the ineffable and/or the ineliminably demonstrative can't be deduced, or even induced or abduced, from a body of impersonal scientific information. (personal communication)

            I daresay that Lycan speaks for many who are sure that Mary learns something, so now we have an explicit rendering of a background presumption of ineffability and an illustration of the role it plays in the argument I call for. Now what about that argument? First of all, nobody could deny that these propositions ventured by Lycan are large theoretical claims, not minimal logical intuitions, or the immediate, unvarnished judgments of experience. *What one knows when one knows what it’s like to experience a blue sensation is ineffable.* I suppose the concept of ineffability being appealed to here would get elaborated along these lines:

It is not the case that there is a string of demonstrative-free sentences of natural language, of any length, that adequately expresses the knowledge of what it is like to experience a blue sensation.

One would like to see that proved. (I’m being ironic. Of all the things one might want to construct a formal theory of, *ineffability* is way down the list, but it might be worthwhile to consider the difficulty of any such undertaking.) Presumably one wants to contrast the ineffability of what it’s like to experience a blue sensation with, say, the ready effability (if I may) of what it’s like to experience a triangle.

Someone who has never seen or touched a triangle can presumably be told in a few well-chosen words just what to expect, and when they experience their first triangle, they should have no difficulty singling it out as such on the basis of the brief description they had been given. They will learn nothing. With blue and red it is otherwise–that, at any rate, is the folk wisdom relied upon by Jackson’s example.  (He wouldn’t have gotten far with a thought experiment about Mary the geometer who was prevented from seeing or touching triangles.) But if what it is like to see triangles can be adequately conveyed in a few dozen words, and what it is like to see Paris by moonlight in May can be adequately conveyed in a few thousand words (an empirical estimate based on the variable success of actual attempts by novelists), are we really so sure that what it is like to see red or blue can’t be conveyed to one who has never seen colors in a few million or billion words? What is it about the experience of red, or blue, that makes this task impossible?  (And don’t just say: they’re *ineffable.* We are enjoined by the extremity of the thought experiment to take this seriously.) Remember, Mary knows *everything* about color that can be learned by physical science, and she presumably has the attention span and powers of comprehension required to handle ten-billion word treatises on what it is like to see red as easily as twenty-five-words-or-less on triangles. Lycan says “at least it’s very tough to put into (noncomparative) words” but this is not a thought experiment about difficulty; it’s a thought experiment about impossibility.  The fact that people find it hard to imagine that any description of what it’s like to see red could do the job is negligible support. Faced with such a formidable task, one does indeed fall back on what Lycan aptly calls the “frustrated demonstrative” but it is a long way from the undeniable claim that it is *very* *tough* to think of ways of characterizing what it is like without resorting to such private demonstratives, to the grand claim that such private demonstratives are strictly speaking ineliminable.  And only absolute ineliminability would carry any weight in an argument against the *possibility* of Mary inferring what would be like for her to see red.  So I stick to my guns. The standard presumption that Mary learns something, that Mary *could* *not* have figured out just what it would be like for her to see colors, is a bit of folk psychology with nothing but tradition–so far–in its favor.

2. You had to be there

            Another unargued intuition exploited by the Mary intuition pump comes in different varieties, all descended inauspiciously from Locke and Hume (think of Hume’s missing shade of blue). This is the idea that the “phenomenality” or “intrinsic phenomenal character” or “greater richness”–whatever it is–cannot be constructed or derived out of lesser ingredients. Only actual experience (of color, for instance) can lead to the knowledge of what that experience is like. Put so boldly, its question-beggingness stands out like a sore thumb, or so I once thought, but apparently not, since versions of it still get articulated. Here are two, drawn from Tye and Lycan:

Now, in the case of knowing via phenomenal concepts, knowing what it is like to undergo a phenomenal state type P demands the capacity to represent the phenomenal content of P under those concepts. But one cannot possess a predicative phenomenal concept unless on has actually undergone token states to which it applies. (Tye, p169)[[4]](http://ase.tufts.edu/cogstud/dennett/papers/RoboMaryfinal.htm" \l "_ftn4" \o ")

As Nagel emphasizes, to know w.i.l., one must either have had the experience oneself, in the first person, from the inside, or been told w.i.l. by someone who has had it and is psychologically very similar to oneself. (Lycan, forthcoming)[[5]](http://ase.tufts.edu/cogstud/dennett/papers/RoboMaryfinal.htm" \l "_ftn5" \o ")

            The role of this presupposition is revealed in the many attempts in the literature to guarantee that Mary doesn’t cheat*,* somehow smuggling the experience of color into her cell.What special care must be taken to prevent Mary from taking surreptitious sips from the well of color?  The blockades erected by Jackson in his original telling have long been recognized as insufficient as they stand. Mary might, for instance, innocently rub her closed eyes one day and create some colored “phosphenes” (try it–I just got a nice deep indigo one right in the middle of my visual field). Or she might use her vast knowledge to engage in some trans-cranial magnetic stimulation of her color-sensitive cortical regions, producing even gaudier effects for her to sort out.  Should a sophisticated alarm system be installed in her brain, to cut short any dream “in color” that she might innocently wander into by happenstance?  Is it in fact possible for a person to dream in color if that person has never seen colors while awake? (Whaddya think? Some might be tempted to respond: “Naw. The colors have to *get in there* through open eyes in order to be available for later use in dreaming.”–that’s the Lockean premise laid bare, and presumably nobody would be seduced by it in such a raw form today.)  If Mary’s color vision system is still intact–a non-trivial empirical assumption, given what is known about the ready reassignment of unused cortical resources in other regards–then she already has “in there” everything she needs to experience color; it just hasn’t been stimulated. (That, at any rate, is the stipulation on which the thought experiment depends, however unrealistic it may be empirically.) A dream could trigger the requisite activity as readily, presumably, as any external stimulus to the open eyes. There are no doubt myriad ways of short-circuiting the standard causal pattern and producing color experience in the absence of external world color.

            More ominously for the prospects of the thought experiment, there are no doubt myriad ways of adjusting the standard causal pattern to produce some state of the brain that is *almost the same as*the sort of state that underlies standard color experience, but that differs in ways that subvert the clarity of the scenario and what it is meant to prove. What started out as a crisp, clean, “intuitive” predicament is being pulled out of shape by the inconvenient complications of science. According to the original thought experiment, it is the subjective, internal *experience of color,*howeverproduced, that is held to be a prerequisite for knowing what it is like to see red, but now that this thesis has lost its naive anchoring in eyes-open-and-awake, it cannot so readily be distinguished from other states of mind that have many of the effects of experiences of color without clearly being experiences of color.  To take the most obvious case, if you right now imagine you are seeing a red rose, do you *thereby* experience red?  (Here is an argument, if a need for one is felt: imagining anything is having an experience, so imagining a red rose is having an experience as of a red rose, which is different from having an experience as of a yellow rose, and the difference must be that in the former case of imagining, you have an experience of red.)  As plausible as this can be made to seem, if it is endorsed, triviality looms for the Mary argument: To know what it is like to experience red is to imagine what it is like and imagine it correctly; but to imagine experiencing red just *is* to experience red, so it follows trivially that you can’t know what it is like to experience red until you have experienced red.

            We are told that Mary in her cell can’t imagine what it’s like to experience red, try as she might.  But suppose she doesn’t accept this limitation and does try her best, cogitating for hours on end, and one day she tells us she just got lucky and succeeded. “Hey,” she says, “I was just day-dreaming, and I  stumbled across what it’s like to see red, and, of course, once I noticed what I was doing I tested my imagination against everything I knew, and I confirmed that I had, indeed, imagined what it’s like to see red!”  Doubting her, we test her by showing her a display of three differently colored circles, and she immediately identifies the red one as red.[[6]](http://ase.tufts.edu/cogstud/dennett/papers/RoboMaryfinal.htm" \l "_ftn6" \o ") What would we conclude?

A. Jackson was wrong; Mary *can* figure out what it’s like to see red in the absence of any experience of red; or

            B. Mary didn’t *figure out* what it is like to see red; she had to resort to (highly intelligent, theory-guided) exercises of *imagining* in order to come to know what it is like to see red. By *imagining*red, she was actually illustrating Jackson’s point, not refuting it. As her example shows, you can’t know what it’s like before you’ve actually experienced what it’s like.

An awkward moment: a simple variation on the tale that clearly refutes it or clearly vindicates it, depending on how you interpret what happened. If B is the only conclusion Jackson intended, then we philosophers have been wasting a lot of time and energy on what appears to be a relatively trivial definitional issue: nothing is going to be allowed to *count* as a state of knowing what it’s like to see red without also counting as an experience of red.

            Before looking more closely at this contretemps, let’s consider one other variation, one I would have thought was the obvious variation for philosophers: Swamp Mary.[[7]](http://ase.tufts.edu/cogstud/dennett/papers/RoboMaryfinal.htm" \l "_ftn7" \o ") Suppressing my gag reflex and my giggle reflex, here she is:

*Swamp Mary:* Just as standard Mary is about to be released from prison, still virginal re colors and aching to experience “the additional and extreme surprise, the unanticipated delight, or the utter amazement that lie in store for her” (G&H, p82), a bolt of lightning rearranges her brain, putting it by Cosmic Coincidence into exactly the brain state she was just about to go into *after* first seeing a red rose. (She is left otherwise unharmed of course; this is a thought experiment.) So when, a few seconds later, she is released, and sees for the first time, a colored thing (that red rose), she says just what she would say on seeing her *second* or *nth* red rose. “Oh yeah, right, a red rose. Been there, done that.”

            Let me try to ensure that the point of this variation is not lost. I am *not* discussing the case in which the bolt of lightning gives Swamp Mary a hallucinatory experience of a red rose. That is, of course, one more “possibility” but it is not the possibility I am introducing. I am supposing instead that the bolt of lightning puts Swamp Mary’s brain into the dispositional state, the competence state, that an experience of a red rose *would have put her brain into* had such an experience (hallucinatory or not) occurred. So, after her Cosmic Accident, Swamp Mary may *think* that she’s seen a red rose, experienced red, been in a token brain state of the type that subserves experiences of red, but she hasn’t. It’s just as if she had.  Maybe she *wrongly remembers or seems to remember* (just like Swampman) having seen a red rose, or maybe, in spite of her lacking any such episodic memories, her competences are otherwise all *as if* she had had such episodes in her past. (After all, you could forget your first color experiences and still have phenomenal concepts, couldn’t you?) *Ex hypothesi* she didn’t have any such experiences, whatever she now thinks; any bogus memories of color were inserted illicitly in her memory box by the lightning bolt.  Hey, [surely] it’s *logically* possible. Swamp Mary is exactly like Mary, an atom-for-atom duplicate of Mary at every moment of her life except for a brief interlude of lightning that performs the accidental (but not supernatural) feat of doing in a flash exactly what Mary’s looking at the rose would do by more normal causal routes. It follows that those who think “that there are certain concepts that . . . . can only be possessed and deployed on the basis of having undergone the relevant conscious experiences oneself” (G&H, speaking of Tye, p65) may be right as a matter of contingent fact, but it is logically possible for one to acquire this enviable ability by accidental means). (These words stick in my throat, but I’m playing the game as best I can.)

3. RoboMary

            We now have two routes to Mary’s post-release knowingness: the Approved Path of “undergoing the relevant conscious experiences oneself” and the logically possible Cosmic Accident Path. The second path is a throw-away, not worth discussing. What *is* worth discussing is a third route to this summit: not a pseudo-miracle but an ascent by good hard work: Mary puts all her scientific knowledge of color to use and *figures out* exactly what it is like to see red (and green, and blue) and hence is not the least bit surprised when she sees her first rose. This third path is hard to imagine, certainly, and as we have just seen, its difficulty is complicated by the threat of a retreat into circularity. It is high time to make the task easier, mounting a positive account that just might convince a few philosophers that they really can imagine it after all. I’m here to help. I will begin with a deliberately simple-minded version, for clarity, and gradually add the complications that the disbelievers insist on. In the spirit of cooperative reverse-engineering, I’m numbering the knobs on my intuition pump, and adding comments on how the knob settings agree or differ from other models of the basic intuition pump

1.RoboMary is a standard Mark 19 robot, except that she was brought on line without color vision; her video cameras are black and white, but everything else in her hardware is equipped for color vision, which is standard in the Mark 19.

Hold everything. Before turning to the interesting bits, I must consider what many will view as a pressing objection:

Robots don’t have color experiences!  Robots don’t have qualia.This scenario isn’t remotely on the same topic as the story of Mary the color scientist.

I suspect that many will want to endorse this objection, but they really must restrain themselves, on pain of begging the question most blatantly. Contemporary materialism–at least in my version of it–cheerfully endorses the assertion that *we* are robots of a sort–made of robots made of robots. Thinking in terms of robots is a useful exercise, since it removes the excuse that we don’t yet know enough about brains to say *just* what is going on that might be relevant, permitting a sort of woolly romanticism about the mysterious powers of brains to cloud our judgment. If materialism is true, it should be possible (“in principle!”) to build a material thing–call it a robot brain–that does what a brain does, and hence instantiates the same theory of experience that we do. Those who rule out my scenario as irrelevant from the outset are not arguing for the falsity of materialism; they are assuming it, and just illustrating that assumption in their version of the Mary story.  That might be interesting as social anthropology, but is unlikely to shed any light on the science of consciousness.[[8]](http://ase.tufts.edu/cogstud/dennett/papers/RoboMaryfinal.htm" \l "_ftn8" \o ")

            Back to knob 1.   Just like Mary,  RoboMary’s *internal* equipment is ‘normal’ for color vision but she is being peripherally prevented–from birth–from getting the appropriate input. RoboMary’s black-and-white cameras stand in nicely for the isolation of human Mary, and we can let her wander at will through the psychophysics and neuroscience journals reading with her black-and-white-camera eyes.

2. While waiting for a pair of color cameras to replace her black-and-white cameras, RoboMary learns everything she can about the color vision of Mark19s. She even brings colored objects into her prison cell along with normally color-sighted Mark 19s and compares their responses–internal and external–to hers.

This was something that Mary could do, of course, only somewhat more tediously–she had to watch black and white TV while conducting all the experiments she needed to get that admirably complete compendium of physical information. This suggests a modest improvement that could be made in Jackson’s original experiment, in which Mary’s eyes are declared normal, and the entire color-blockade has to be accomplished with prison walls, confiscation of mirrors, white gloves, etc. As various commentators have observed, such a world would still be an ample source of chromatic input–shadows, and the like, not to mention the different shades of “white”.  It would have been a lot cleaner for Jackson’s original telling if he had stipulated that Mary had a pair of camcorders with black-and-white eyepieces strapped over her eyes, peering at the world all her life like somebody videotaping her vacation in Europe. (Or, slightly more science-fictionally,  he might have imagined Mary not imprisoned but with “filters” implanted on her optic nerves, permitting only black-and-white signals through.)

3. She learns all about the million-shade color-coding system that is shared by all Mark19s.

We don’t know that human beings share the same color-coding system. Probably they don’t, but this is just a complication we can leave out; if Mary knows *everything*, she knows all the variations of human color-coding, including her own.

4. Using her vast knowledge, she writes some code that enables her to colorize the input from her black and white cameras (à la Ted Turner’s cable network) according to voluminous data she gathers about what colors things in the world are, and how Mark19s normally encode these. So now when she looks with her black-and-white cameras at a ripe banana, she can first see it in black and white, as pale gray, and then imagine it as yellow (or any other color) by just engaging her colorizing prosthesis, which can swiftly look up the standard ripe-banana color-number-profile and digitally insert it in each frame in all the right pixels. After a while, she decides to leave the prosthesis turned on all the time, automatically imagining the colors of things as they come into focus in her black and white camera eyes.

Isn’t this simply the robot version of phosphenes and trans-cranial magnetic stimulation–forbidden ways of getting color experience into RoboMary? Or is it rather a way of dramatizing the immense knowledge of color “physiology” that RoboMary, like Mary, enjoys? What is either of them allowed to do with their knowledge?  Is this a cheat or isn’t it?

Let’s turn the knob both ways, and see what happens. In the first, and simplest, setting, we declare that just as Mary is entitled to use her imagination in any way she likes in her efforts to come up with an anticipation of what it’s going to be like to see colors,  RoboMary is entitled to use her imagination, and that is just what she is doing–after all, no hardware additions are involved: she is just considering, by stipulation, what it might be like to see color under various conditions. (We can suppose she goes to the trouble of considering dozens of variant colorization codings, so she has entertained many different hypotheses about what it is like to see red, etc., and settled, defeasibly, on the one she thinks is best.)

5. She wonders if the ersatz coloring scheme she’s installed in herself is high fidelity.  So during her research and development phase, she checks the numbers in her registers (the registers that transiently store the information about the colors of the things in front of her cameras) with the numbers in the same registers of other Mark 19s  looking at the same objects with their color camera eyes, and makes adjustments when necessary, gradually building up a good version of normal Mark 19 color vision.

In the case of RoboMary it is obvious what sorts of use she can make of her knowledge about color and color vision in Mark 19s. It is far from obvious, of course, how human Mary could make use of *her* knowledge. But that just shows how treacherous the original intuition pump is; it discourages us from even trying to imagine the task facing Mary if she wants to figure out what it is like to see red.

6. The big day arrives. When she finally gets her color cameras installed, and disables her colorizing software, and opens her eyes, she notices . . . . nothing. In fact, she has to check to make sure she has the color cameras installed. She has learned nothing. She already knew exactly what it would be like for her to see colors just the way other Mark 19s do.

4. Locked RoboMary

Too easy! Now let’s turn the knob and consider the way RoboMary must proceed if she is prohibited from tampering with her color-experience registers. I don’t know how Mary could be crisply rendered incapable of using her knowledge to put her own brain into the relevant imaginative and experiential states, but I can easily describe the software that will prevent RoboMary from doing it. In order to prevent this sort of self-stimulation skullduggery (if that is what it is), we arrange for RoboMary’s color-vision system–the array of registers that transiently hold the codes for each pixel in Mary’s visual field, whether seen or imagined–to be restricted to gray-scale values. This is simple: We arrange to code the gray-scale values (white through many shades of gray to black) with numbers below a thousand, let’s say, and simply filter out (by subtraction) any values for chromatic shades in the million-shade subjective spectrum of Mark 19s–and we put unbreakable security on this sub-routine. Try as she might, RoboMary can’t jigger her “brain” into any of the states of normal Mark 19 color vision *or imagination*. She has all her hard-won knowledge of that system of color vision, but she can’t use it to adjust her own hardware so that it matches that of her conspecifics. Her color-representing hardware is disabled.

            This doesn’t faze her for a minute, however. Using a few terabytes of spare (undedicated) RAM, she builds a model of herself and *from the outside*, *just as she would if she were building a model of some other being’s color vision*, she figures out just how she would react in every possible color situation.

            I find that people have trouble imagining just how intimate and extensive this “third-person” knowledge would be, so let’s indulge in a few illustrative details, to help furnish our imaginations. She obtains a ripe tomato and plunks it down in front of her black-and-white-cameras, obtaining some middling gray-scale values, which lead her into a variety of sequel states. She automatically does the usual “shape from shading” algorithm, obtaining normal convictions about the bulginess and so forth, and visually guided palpation gives her lots of convictions about its softness. She consults an encyclopedia about the normal color range of tomatoes, and she knows that these gray-scales in these lighting conditions are consistent with redness, but of course nothing comes to her directly about color, since she has black-and-white cameras, and moreover, she can’t use her book-learning to adjust these values, since her color system is locked.  So, as advertised,  she can’t put herself directly into the *red-tomato-experiencing* state. She looks at the (gray-appearing) tomato and reacts however she does, resulting in, say, thousands of temporary settings of her cognitive machinery. Call that voluminous state of her total response to the locked gray-tomato-viewing *state A*. This is a state of her knowing what it is like for her to see a gray tomato. Then she compares state A with the state that her model of herself goes into. Her model isn’t locked; it readily goes into the state that any normal Mark 19 would go into when seeing a red tomato. And, since this is her model of herself, it then goes into *state B*, the state she would have gone into if her color system hadn’t been locked.  RoboMary notes all the differences between state A, the state she was thrown into by her locked color system, and state B, the state she would have been thrown into had her color system not been locked, and–being such a clever, indefatigable and nearly omniscient being–makes all the necessary adjustments and *puts herself into state B.*State B is, by definition, *not* an illicit state of color-experience; it is the state that such an illicit state of color-experience normally causes (in a being just exactly like her). But now she can know just what it is like for her to see a red tomato, because she has managed to put herself into just such a dispositional state–this is of course the hard-work analog of the miraculous feat wrought by the Cosmic Accident of the lightning bolt in the case of Swamp Mary.

            Her epistemic situation when she has completed this Vast (Dennett, 1995, p109) but not infinite labor is indistinguishable from her epistemic situation in the case in which we allowed her to colorize her actual input–and it had been conceded that in that epistemic situation she had known what it is like to see red, but the case was thrown out for cheating. So there are no surprises for her when her color system is unlocked and she’s given color cameras. In fact, when she completes her model of herself, down to the very last detail, she can arrange for it to take over for her locked onboard color system, a spare color system she can use much as Dennett used his spare computer brain in “Where am I?” (in Dennett, 1978). Remember: RoboMary knows all the physical facts, and that’s a lot.

            Objection: RoboMary can’t put herself into state B, the state her model is driven into by its unlocked color system, because that state involves the wielding of what Tye calls “phenomenal concepts” and these are strictly parasitic on actual phenomenal experiences, which they quote or reproduce, in effect, when they are exploited in such demonstrative thoughts as “*that* is what red looks like.” Part of having the competence that comes (normally) from experience is being able later to use demonstratives with internal referents of this sort.

            Oh really? Why can’t RoboMary form demonstratives that allude to the relevant states of her model, instead of her own locked color system? And why wouldn’t they be just as good? Because they wouldn’t have that extra *je ne sais quoi*?  But that is just what has not been shown to exist. In the case of RoboMary the temptation to posit a rather magical extra property that adheres somehow to her entering into these color-system states (which are basically just numbers in registers, after all) is weak. The temptation should be resisted in the case of Mary, too. It has no legitimate business to do, and tends to distort the imagination covertly.

            Objection (thanks to the editors of this volume): For RoboMary to self-program herself into state B is cheating just as much as for her to self-program herself into the “experiencing red” state. What matters is whether Mary (or RoboMary) can *deduce* what it’s like to see red from her complete physical knowledge, not whether one could use one’s physical knowledge in some way or other to acquire knowledge of what it’s like to see in color.

            I just don’t see that this is what matters. So far as I can see, this objection presupposes an improbable and extravagant distinction between (pure?) deduction and other varieties of knowledgeable self-enlightenment. I didn’t describe RoboMary as “self-programming” herself; I said she “notes all the differences between state A, the state she was thrown into by her locked color system, and state B, the state she would have been thrown into had her color system not been locked, and–being such a clever, indefatigable and nearly omniscient being–makes all the necessary adjustments and *puts herself into state B.*”  If I use my knowledge to imagine myself into your epistemic shoes in some regard, is this “self-programming”? And if so, what is the import of this characterization for the knowledge argument? Consider Rosemary, another of Mary’s daughters, who is entirely normal and free to move around the colored world, and is otherwise her mother’s equal in physical knowledge of color.  Rosemary has a hard time imagining her mother’s epistemic predicament. What must it be like, she wonders, not yet to know what it is like to see red?  She is burdened, it seems, with *too* *much* knowledge (cf. my example of the newly discovered Bach Cantata in *Consciousness Explained*, 1991, p388).  This is, presumably, a psychological impediment to her imagination, but not an epistemological lack.

            I take the example of RoboMary to shift the burden of proof. Thin materialism, the view that Mary, in her well nigh unimaginable circumstances, would not be surprised after all, has a lot to be said for it. Enough, surely, to undermine the blithe confidence with which philosophers have presumed otherwise.

            A closing observation: I find that some philosophers think that my whole approach to qualia is not playing fair. I don’t respect the standard rules of philosophical thought experiments. “But Dan, your view is so *counterintuitive*!” No kidding. That’s the whole point. Of course it is counterintuitive. Nowhere is it written that the true materialist theory of consciousness should be blandly intuitive. I have all along insisted that it may be *very* counterintuitive. That’s the trouble with “pure” philosophical method here. It has no resources for developing, or even taking seriously, counterintuitive theories, but since it is a very good bet that the true materialist theory of consciousness will be highly counterintuitive (like the Copernican theory--at least at first), this means that “pure” philosophy must just concede impotence and retreat into conservative conceptual anthropology until the advance of science puts it out of its misery. Philosophers have a choice: they can play games with folk concepts (ordinary language philosophy lives on, as a kind of aprioristic social anthropology) or they can take seriously the claim that some of these folk concepts are illusion-generators. The way to take that prospect seriously is to *consider* theories that propose revisions to those concepts.

References:

Dennett, Daniel, 1978, *Brainstorms: Philosophical Essays on Mind and Psychology*, Cambridge, MA: MIT Press/A Bradford Book.

Dennett, Daniel, 1991, *Consciousness Explained*, Boston: Little Brown.

Dennett, Daniel, 1994, “Get Real,” reply to my critics in  *Philosophical Topics*, **22**,  pp. 505‑568

Dennett, Daniel, 1995, *Darwin’s Dangerous Idea*, New York: Simon & Schuster.

Dennett, Daniel, 1996, “Cow‑sharks, Magnets, and Swampman,” in *Mind & Language*, vol. 11, no. 1, pp. 76‑77.

Graham, George, and Terence Horgan, 2000, “Mary Mary Quite Contrary,” *Philosophical Studies*, **99**, pp59-87.

Hofstadter, Douglas, 1981,  “Reflections,” in Hofstadter and Dennett, ed., *The Mind’s I*, New York: Basic Books, 1981, p375.

Lycan, William, 1996, *Consciousness and Experience*, Cambridge, MA: MIT Press.

Lycan, William, forthcoming, “Perspectival Representation and the Knowledge Argument,” in Q. Smith and A.Jokic (eds.), *Consciousness: New Philosophical Essays,*Oxford U.P.

McConnell, Jeff, 1994, “In Defense of the Knowledge Argument,” *Philosophical Topics*, **22**, pp157-897.

McGeer, Victoria, forthcoming, “The Trouble With Mary,” WHERE?

Robinson, Howard, 1993, “Dennett on the Knowledge Argument,” *Analysis*, **53**, pp174-77,

Tye, Michael, 1995, *Ten Problems of Consciousness*, Cambridge, MA: MIT Press.

[[1]](http://ase.tufts.edu/cogstud/dennett/papers/RoboMaryfinal.htm" \l "_ftnref1" \o ")I am grateful to Diana Raffman, Bill Lycan, Victoria McGeer and my students for many discussions, on email and in person, on the ins and outs of this argument.

[[2]](http://ase.tufts.edu/cogstud/dennett/papers/RoboMaryfinal.htm" \l "_ftnref2" \o ")For instance, Howard Robinson (1993) supposes that I am illicitly helping myself to the premise that Mary knows “every particular physical thing that is going on” (p175), but my claim does not at all depend on such a strong claim, as will be clear from the variations I develop here.

[[3]](http://ase.tufts.edu/cogstud/dennett/papers/RoboMaryfinal.htm" \l "_ftnref3" \o ")Robinson (1993) also claims that I beg the question by not honoring a distinction he declares to exist between knowing *“what one would say and how one would react”* and knowing “what it is like.”  If there is such a distinction, it has not yet been articulated and defended, by Robinson or anybody else, so far as I know.  If Mary knows *everything* about what she would say and how she would react, it is far from clear that she wouldn’t know what it would be like.

[[4]](http://ase.tufts.edu/cogstud/dennett/papers/RoboMaryfinal.htm" \l "_ftnref4" \o ")Earlier Tye had noted simply that “possessing the phenomenal concept *red* requires that one have experienced red . . . .  possession consists (very roughly) in having available a state that has a causal history that links it with the relevant experiences . . . . ” (p167)

[[5]](http://ase.tufts.edu/cogstud/dennett/papers/RoboMaryfinal.htm" \l "_ftnref5" \o ")Lycan’s second alternative is a startling concession, unremarked upon. If one can be told, why can Mary not simply be told, by standard color-enjoyers psychologically very similar to her? I don't think Nagel allows this, and if Lycan does, the game is over. Thin materialism wins. In the context, Lycan is imagining the variation in which Mary is born colorblind, which explains the lapse: those who could tell her are not ‘psychologically very similar’ to her.

[[6]](http://ase.tufts.edu/cogstud/dennett/papers/RoboMaryfinal.htm" \l "_ftnref6" \o ")I find that this example nicely focusses the attention of students. I show my students a display that contains three brightly-colored and unlabeled disks of red, green and blue: “Are you claiming that Mary *could not tell*, on first seeing this display, which disk was red?” They are inclined to insist that she could not, but what would they make of it, I then ask, if she could do it? Would it have to be “magic” or “cheating” or could it just possibly be the result of good hard work on her part? They then see that the impossibility claim is in fact much more dubious than Jackson’s “it seems just obvious” line invites us to suppose.

[[7]](http://ase.tufts.edu/cogstud/dennett/papers/RoboMaryfinal.htm" \l "_ftnref7" \o ")Gabriel Love suggested this hybrid to me, and I think he was inspired by considering another: In “The Trouble With Mary,” Victoria McGeer (forthcoming) has written persuasively and elegantly on what happens when you hybridize Mary and zombies. I have expressed my distrust of all such thought experiments in Dennett (1994); excerpted (with revisions) in Dennett, (1996).

[[8]](http://ase.tufts.edu/cogstud/dennett/papers/RoboMaryfinal.htm" \l "_ftnref8" \o ")It has been drawn to my attention by the editors that a “type-identity theorist who believes conscious experiences are essentially neural” could balk at my robot scenario, on grounds that it was impossible. Officially, I suppose, such a theorist counts as a materialist, and so would a theorist who views blood-pumping as “essentially cardial” and declares artificial hearts to be impossible, but nobody has yet offered a remotely plausible reason for believing any such doctrine. Such mysterian or vitalistic versions of materialism deserve to be ignored until someone has made a positive case for them.