

Bounded Rationality, Evolutionary Design Constraints, and a Neuroscientist Named ‘Mary’

Gregory R. Mulhauser
Cognitive & Evolving Systems Research
Centre for Cognition, Communication and Computation
British Telecom Laboratories
Martlesham Heath, England IP5 3RE
www.labs.bt.com/people/mulhaug
scarab@info.bt.co.uk

Abstract. Arguing that the constraints on human cognitive transitions which bear most urgently on the project of naturalising the study of mind are set by evolution and underlying physics, not logical possibility, this paper begins by showing that physically instantiated cognizers can have no more computational power than finite automata. It finishes by suggesting that far from constituting an argument *against* physicalism, Jackson’s (1982) famous story about colour-deprived neuroscientist Mary runs just the same even if we *assume* physicalism. The abiding failure to identify it as an observation of the *consequences* of physicalism stems from a reluctance to ‘take physicalism seriously’.

Word count for abstract: 100

Total word count (including footnotes & references, excluding abstract): 3963

Word count excluding footnotes, references & abstract: 3399

Bounded Rationality, Evolutionary Design Constraints, and a Neuroscientist Named ‘Mary’

Gregory R. Mulhauser
Cognitive & Evolving Systems Research
Centre for Cognition, Communication and Computation
British Telecom Laboratories
Martlesham Heath, England IP5 3RE
www.labs.bt.com/people/mulhaug
scarab@info.bt.co.uk

Abstract. Arguing that the constraints on human cognitive transitions which bear most urgently on the project of naturalising the study of mind are set by evolution and underlying physics, not logical possibility, this paper begins by showing that physically instantiated cognizers can have no more computational power than finite automata. It finishes by suggesting that far from constituting an argument *against* physicalism, Jackson’s (1982) famous story about colour-deprived neuroscientist Mary runs just the same even if we *assume* physicalism. The abiding failure to identify it as an observation of the *consequences* of physicalism stems from a reluctance to ‘take physicalism seriously’.

1. Introduction

At least three considerations motivate the study of constraints on cognitive agents’ reasoning ability. First, in business and the social sciences, some understanding of the real world (as distinct from ‘in principle’) constraints on cognitive agents is indispensable for developing models of both individual and group behaviour—including economic models, management models, and even computational models of artificially intelligent agents. Different constraint types and different values of those constraints may result in strong quantitative as well as qualitative differences in behaviour. Second, an understanding of constraints helps shape and guide the naturalist programme itself: a naturalistic account of the mind must ultimately identify the natural constraints on cognitive ability and explain cognition within those constraints. If, for instance, one believes that the limits on human agents’ cognitive processes are all and only those set by logical possibility, one will likely consider a different range of scientific theories than will someone who believes there are other, more stringent constraints. Finally and most simply, knowing our limits helps us to grasp who and what we are as cognizers.

In addition to logical possibility, I believe at least two varieties of constraints limit human cognitive agents’ abilities. After many years of neglect, increasing attention has more recently come to bear on the first of these two kinds: limits imposed by *computational tractability*. For instance, balancing the agent’s need for deductive power with real constraints on time and resources, Cherniak observes that “not only is acceptance of a metatheoretically adequate deductive system not transcendently indispensable for an agent’s rationality, but in important cases it is inadvisable and perhaps even incompatible with that rationality” (1984, p. 755). It is with a second variety that this paper is principally concerned: namely, the limits imposed

on agents by virtue of their physical structure and the factors, such as development and phylogenetic history, which shape that structure. The discussion about Jackson's neuroscientist in section 4 is intended to underscore the sorts of difficulties which may arise more generally when limits of this second type receive inadequate attention, particularly as they bear on the second motivation mentioned above.

2. It's Just Obvious, Isn't It?

The basic line of thought is seductively simple:

- "In principle, I can count as high as I like: 1, 2, 3, ... 48,536,012, ..."
- "I could construct infinitely many well-formed English sentences, given enough time."
- "Given any formal system, I can always step into a higher level system and prove meta-theorems about the first."

Few warm to the idea that their capabilities might, computationally speaking, be limited to those of a 'mere' finite automaton. Much more attractive is the fantasy that we're capable of universal computation and are at least as powerful as Universal Turing Machines, if not more so. And the capabilities above, all of which require more power than that of a finite automaton, seem so obviously part of the human repertoire.

While many theorists might just be happy with the obviousness of it all, some (Lucas 1961, Penrose 1989, Bringsjord 1992) have even taken a stab at principled arguments purporting to show that the mathematical or creative capabilities of human beings categorically exceed those of any formal system or classical mechanical device. Others, like Kripke (1980, 1982), suggest that not even the meanings of our concepts can be captured by formal systems, while Dummett, in 'The Philosophical Significance of Gödel's Theorem' (reprinted in Dummett 1978), argues that our notion of 'proof' certainly cannot be captured formally, because any putative account will always miss out some higher level meta-proofs which that particular formal account cannot accommodate.

Probably each of these theorists would roll their eyes and admit readily that *of course* the real world constrains the kinds of 'in principle' claims suggested above: my heart will stop beating before I count to infinity, I'll utter only a finite number of sentences in my lifetime, and some formal systems might just be so huge that I could work for my entire career without generating any meta-theorems about them. Such constraints are so plain and boring that perhaps everyone just takes them for granted. But what seems to have captured these theorists' attention is the notion that *in principle*, I could still perform each of those feats (if only my heart never stopped beating, etc). *Prima facie*, this seems as harmless as the observation that no Universal Turing Machine which computes a function and halts ever *actually* uses an infinite amount of tape, because the point is that the machine's fundamental architecture guarantees universal power: it has the inherent power, in principle, to use infinitely much tape, were it available.

2.1 Black Holes in the Ointment

But it turns out that *any* physical system which can be enclosed within a manifold of finite surface area can exist in only a finite number of distinguishable configurations. This fundamental result from research on the

thermodynamics of black holes¹ (Bekenstein 1981) appears to imply straightforwardly that the computational power of *any* such real physical system is bounded by that of finite automata. (Note that I am *not* claiming that all such systems are best modelled by finite automata, or that they directly implement them, or that they are formal systems at all—but merely that their computational capabilities are *bounded* by those of finite automata.)

In other words, if real biological cognizers such as human beings are instantiated wholly physically, then it would appear that no such entities can have universal computational power. *Of course* we can object that if only there were no such physical barriers, then human beings could be as powerful as Universal Turing Machines. But this is little different from saying that if only finite automata weren't finite, then they, too, could be as powerful as Universal Turing Machines. (Away with those meddlesome computational classes: everything could be infinite if only it weren't finite!)

That is, it isn't saying much.

In the absence of some *independent* reason for thinking either that 1) Bekenstein has his physics wrong or 2) human cognition is not implemented physically, the above reasoning should be enough to deflate the notion that we have anything beyond the power of finite automata. By way of 'independent reason', we would require something beyond the intuitions behind each of the 'obvious' capabilities above. After all, each time I've been behind the wheel of a running car (on a level surface, with good traction, with transmission engaged, etc.), pressing the accelerator brought it about that the car went faster. But in the face of our knowledge of basic physics, this is certainly not enough to make us believe that *every* time I press the accelerator, the car will go faster. Stepping on the accelerator works only until the car reaches its fundamental limits. Likewise, the fact that every time someone has asked me to add 1 to a given number, I've been able to do it, provides no rationale whatsoever for believing that I am capable of adding 1 to any old number at all.

Before leaving the topic, there is one consolation of sorts for anyone unhappy about missing out on universal computational power. Namely, no finite series of measurements could ever reject the hypothesis that a real physical system was only computationally equivalent to a finite automaton as opposed to some variety of universal computer. This suggests that the hypothesis that a real physical system such as a human cognizer has universal computational power is not even empirically meaningful. Only theoretical considerations can settle the issue, and the conclusion thus bears only on the second and third of the reasons given in section 1 for caring about bounds on cognitive ability.

3. Take Physicalism Seriously!

While Chalmers's (1996) admonition to *take consciousness seriously* has become a clarion call in some quarters of philosophy of mind and theoretical cognitive science, I suggest another: take *physicalism*

¹ Why black holes? Because as Bekenstein (1981, p. 287) observes, "black holes have the maximum entropy for given mass and size which is allowed by quantum theory and general relativity": thus, any upper bound on the number of internal configurations for a black hole applies with equal force to *every other physical system*.

seriously! By this I mean that if one is to be a physicalist, then one should pay careful attention to what being physically instantiated implies about cognizers. This might include, as suggested above, ensuring that we don't begin with implicitly contradictory assumptions—such as that we can simultaneously be finite physical entities and be capable of performing computational feats which cannot be accomplished by any finite physical entity. It also means viewing our own capacities for logical thought, for conscious experience, and for thought about that conscious experience, in terms of the physical substrate which implements those capacities.

However we choose to 'precisify' physicalism, hopefully most will agree that it incorporates at least the following notion: that when a cognizer enters a new cognitive or phenomenal state (say, of grasping the meaning of a proposition, or of experiencing a butterfly in the left visual field), this requires that the cognizer's microphysical substrate also enters a new state. More to the point, any given psychological or experiential transition requires a microphysical transition to some state in a specific class of microphysical states 'associated' with the new state. This is just a long-winded way of saying that cognitive and phenomenal state transitions supervene on microphysical state transitions, with the meaning of the word 'associated' determined by the character of the logical relationship underlying the supervenience relation (e.g., *a priori* or *a posteriori* intension, etc.). For now, we may set aside concerns about modality, rigidification of designators referring to mental states, and the derivability of supervening propositions from the supervenience base; all that is required for present purposes is the notion that a given transition at the level of cognition or experience requires some underlying microphysical transition. (For a more careful treatment of the structure of supervenience relations in this context, see Mulhauser 1998.)

This suggests that, in the case of biological creatures, any transition of 'reasoning' or 'logic' at a cognitive level is subserved by a physical transition at some level of biological wetware—and that whatever may constrain such physical transitions in our biological wetware may constrain the supervening transitions such as reasoning. Likewise, for whatever cognitive or experiential transitions we find do not occur for whatever reason at the cognitive level, there will be an analogous story about transitions failing to occur in microphysical state space. This point is very simple and altogether independent of the logical structure of the relationships² between explanations which might be given at the levels of agents' cognition and of microphysics. I am simply observing that if a transition does not occur at a supervening level, then the corresponding transition does not occur in the supervenience base (and *vice versa*).

To the extent that constraints on our reasoning may be set by the laws of physics and by the processes of evolution by natural selection which shaped our physical configurations through phylogenetic time, this suggests a useful area of research complementing the central aims of evolutionary psychology (Tooby & Cosmides 1992; see Mitchell 1998 for a recent review). While evolutionary psychology aims principally to

² In particular, the point is independent of worries about explanatory exclusion (Kim 1989). Also, it is obviously true that if the mapping from microphysical states to inclusive psychological states is many-to-one, then transitions may occur at the former level in the absence of transitions at the latter level, and constraints on the former may or may not translate into constraints on the latter; nothing here contradicts this.

discover the positive adaptive explanations behind architectural features of human psychology, we might also look to explain the *bounds* on that architecture—on our rationality and general cognitive abilities—by attending to the physical constraints which specifically arise as byproducts of positive adaptation of other characteristics.

Picking very fluffy examples just to communicate the flavour of what I have in mind, consider an evolutionary explanation of characteristics of human vision. Just as we might give a positive explanation of why the lens of the eye is circular in one plane in terms of the adaptive value of focusing light onto the retina with minimal distortion, so too might we explain why humans cannot see in the infrared spectrum not just in terms of weak selective pressures for infrared vision but also in terms of the architectural side effects of positive adaptations to the selective pressure to react to a certain range of higher wavelengths. Perhaps the best adaptation for visible wavelengths was just architecturally incompatible with infrared vision. Similarly, not only might we find an evolutionary story accounting for humans' rich capacity for associative memory, but we might also find an evolutionary story explaining why humans lack a knack for 128-bit precision binary arithmetic not just in terms of an absence of selective pressure for performing high precision binary arithmetic but *also* in terms of the architectural side effects of *positive* adaptations for other traits which *were* under selective pressure. In other words, the study of constraints is in part a study of the evolutionary spandrels (Gould & Lewontin 1978; or 'pedentives', as Dennett 1995 prefers) of cognition. (See Polger & Flanagan forthcoming for the view that consciousness itself is an evolutionary spandrel, or etiological epiphenomenon.)

So, to rehearse the point, taking physicalism seriously means paying attention to the basic physical boundary conditions which constrain the capabilities of physically instantiated cognizers. It means paying attention to the evolutionary story behind cognizers' physical substrates and thinking not only in terms of the logical structure of cognitive transitions but also in terms of the underlying transitions through state space which subserve them. While the discussion in the following section unfortunately provides no such tidy evolutionary explanation as alluded to above, it offers a different perspective on Frank Jackson's (1982) Mary which I hope will highlight the potential pitfalls of failing to take physicalism seriously.

4. Something About Mary

In a thought experiment whose mystery has so far shown few signs of waning as it nears its third decade, Jackson invites us to consider a neuroscientist who has never actually seen the colour red but who knows all there is to know about the physical details of colour vision. She knows what wavelengths of light are reflected by red objects, she knows how those wavelengths affect the human retina, and how signals from the retina ascend via the optic nerve, LGN, and so forth (or superior colliculus, etc.) into the rest of the brain. She knows all that happens in the brain of a subject experiencing a visual sensation of the colour red. Yet, Jackson plausibly suggests, Mary *still* learns something new when she herself sets eyes upon a red object for the first time. Thus, tugs the tempting conclusion, not all that there is to know about seeing the colour red can be physical—after all, Mary already knew all that, *ex hypothesi*—so physicalism must be false.

(As an aside, it's worth noting that physicalism *per se* needn't require that all propositions about phenomenal experience be *formally derivable* from propositions about physical facts. Physicalism as mildly 'precisified' above requires only that cognitive transitions supervene in some sense on microphysical transitions. It takes a great deal more work to show formal derivability for a given choice of specific supervenience relation—as distinct from the more popular ploy of simply *assuming* that a logic which is both complete and consistent will fit the bill.³ Even were the basic logical footwork to be completed, showing that derivation is *computationally tractable* for Mary is another task altogether. As far as I am aware, no one has ever done this. So for now, the jury ought really to be out on whether Mary learns something new upon seeing a red rose which she couldn't have derived from what she already knew. But I won't pursue that line: Jackson's story *isn't about formal derivability*.)

Jackson's story is about our intuition that even after learning everything physical there was to know about the colour red and colour vision, Mary might still learn something new upon seeing the colour for the first time. I think intuition is right: Mary learns something new.

But, trying to take physicalism seriously, consider the story from the vantage point of underlying physical state transitions. Let us *assume* physicalism and run through the story. For Mary to see red is for her physical substrate to be in some state which is a member of the class of points in state space which subserve the experience of seeing red. Likewise, for Mary to 'know what it is like' to see red is for her to be in a state in some other class, which overlaps to a significant degree with the first (and which *perhaps* contains the first as a proper subclass, depending upon whether experiencing red automatically yields knowledge of what it is like to see red). And for Mary to know physiological or other physical facts about colour vision is for her to be in still another class, comprising those points in state space which subserve whatever cognitive attributes define knowing about colour vision.

We may assume that there does exist a state in which Mary knows what it is like to see red, and there does exist a state in which Mary knows all about colour vision but not what it is like to see red; the thought experiment is about the challenge of getting from one to the other.⁴ It is about the failure of that transition to occur until Mary actually sees a red object and thereby learns what it is like. The reader is invited to think that physicalism means Mary's powers of logic should be able to effect just such a transition, and that the failure of the transition requires that physicalism be rejected.

³ As is well known, in a stronger logic, it *may* be the case that formal derivations of all the relevant propositions only become available at the cost of rendering the logic inconsistent. Simply assuming to the contrary isn't even taking *logic* seriously.

⁴ On an alternative reading, the class of points subserving states of knowing what it is like to see red properly contains the class subserving states of knowing all the physical facts about colour vision. On this view, anyone who knows all the physical facts should *automatically* know what it is like, without even thinking about it. But not even the strongest supervenience relation between microphysics and phenomenal experience, one grounded in *a priori* intension (i.e., intension fixing extension for whatever world is actual, as distinct from intension fixing extension for counterfactual worlds, *given* actual world extension), requires *that* sort of view.

Nothing could be further from the truth.

The failure of such a transition is an accident of the way Mary is physically constructed. For proof of this fact, we need only consider a scenario where Mary is equipped with some hypothetical artificial retinal stimulator (or optic nerve stimulator, cortical stimulator, etc.). The idea of artificial nerve stimulators has a long and suspicious history in such thought experiments, but the point is nonetheless an important one: by using her complete physical information about colour vision to modify her brain state directly with an artificial stimulator, Mary can effect the physical state transition which subserves a cognitive transition from a state of knowing all about colour vision to a state of actually seeing the colour red. In other words, by using her knowledge of what sort of brain state would elicit a sensation of red, together with an artificial aid to stimulate (or inhibit) her brain, she can in principle bring about just the transition which does not normally occur without artificial assistance or an actual red object to observe. However, it is purely an accident of physiology that she must take recourse to an artificial stimulator: there is nothing at all incoherent about supposing that Mary could have been born with some unusual pathways which allowed her to twiddle nerves in her ascending visual pathways just by thinking about it.

Finally, this is the kicker: *even if* Mary's powers of logic could have led her to derive, from her complete understanding of colour vision, every detail of a Mary-experiencing-red brain state (or a Mary-knowing-what-it-is-like-to-see-red state), the state of Mary-experiencing-red *is not the same state* as that of Mary-knowing-the-brain-state-of-Mary-experiencing-red. *More generally, knowing a complete description of a brain state X requires being in a different state than X.*⁵ In other words, Mary is simply not *physically* equipped to enter given psychological or experiential states—such as seeing red, knowing what it is like to see red, or happiness at gaining tenure—just by reading and understanding descriptions of those brain states on which they supervene. Crucially, this has nothing whatsoever to do with the failure of physicalism but is indeed an *implication* of physicalism taken together with the basic physical boundary constraints set by our brain structure.

There is a great deal more to be said on this issue, and in particular about the perspectival nature of experience. (Mulhauser 1998) Not surprisingly, matters such as this are rarely so simple as they might at first appear. But I believe the basic conclusion itself is inescapable: Jackson's thought experiment about Mary, which has for nearly twenty years cast a shadow of doubt over the entire naturalist project of understanding the mind via existing natural laws governing the physical world, is not what it seems. It is in fact *an observation of one of the consequences of the truth of physicalism*. And I believe the principal reason this point seems to have eluded commentators for so long is an abiding reluctance to take physicalism seriously—constraints and all.

⁵ OK, there is probably one class of exceptions. The state of understanding a description of a state may be the same as the state itself when the description is actually *of* the state of understanding a description of the state. However, this class of pathologically weird fixed points in the mapping between microphysics and cognition does not compromise the point being made.

References

- Barkow, J.H.; L. Cosmides, and J. Tooby, eds. (1992) *The Adapted Mind*. Oxford: Oxford University Press.
- Bekenstein, J.D. (1981) 'Universal Upper Bound on the Entropy-to-Energy Ratio for Bounded Systems', *Physical Review D* 23(2): 287-98.
- Bringsjord, S. (1992) *What Robots Can and Can't Be*. Dordrecht: Kluwer Academic.
- Cherniak, C. (1984) 'Computational Complexity and the Universal Acceptance of Logic', *Journal of Philosophy* 81: 739-55.
- Chalmers, D.J. (1996) *The Conscious Mind*. Oxford: Oxford University Press.
- Dennett, D.C. (1995) *Darwin's Dangerous Idea*. London: Simon & Schuster.
- Dummett, M. (1978) *Truth and Other Enigmas*. London: Duckworth.
- Gould, S. J. & R. Lewontin. (1978) 'The spandrels of San Marco and the Panglossian paradigm: A critique of the adaptationist program', *Proceedings of the Royal Society of London* B205: 581-98.
- Jackson, F. (1982) 'Epiphenomenal Qualia', *Philosophical Quarterly* 32: 127-36.
- Kim, J. (1989) 'Mechanism, Purpose, and Explanatory Exclusion', *Philosophical Perspectives* 3: 77-108.
- Lucas, J.R. (1961) 'Minds, Machines and Gödel', *Philosophy* 36: 112-27.
- Mitchell, M. (1998) 'Can Evolution Explain How the Mind Works?', Santa Fe Institute report 98-11-105. To appear in *Complexity*.
- Mulhauser, G.R. (1998) *Mind Out of Matter: Topics in the Physical Foundations of Consciousness and Cognition*. Dordrecht: Kluwer Academic.
- Mulhauser, G.R., ed. (forthcoming) *Evolving Consciousness*. To appear from John Benjamins.
- Penrose, R. (1989) *The Emperor's New Mind*. Oxford: Oxford University Press.
- Polger, T. & O. Flanagan (forthcoming) 'Consciousness, Adaptation and Epiphenomenalism', in Mulhauser (forthcoming).
- Tooby, J. & L. Cosmides (1992) 'The Psychological Foundations of Culture', in Barkow, et al. (1992).