Electronic version of an article published as:
Int. J. Mach. Conscious. Vol. **05**, No. 01, 2013, pp. 59-74
doi: 10.1142/S1793843013400052 © World Scientific Publishing Co.
http://www.worldscientific.com/doi/abs/10.1142/S1793843013400052

SAFE/MORAL AUTOPOIESIS & CONSCIOUSNESS

MARK R. WASER

Digital Wisdom Institute
610 Kearney Court, SW
Vienna, VA 22180, USA
mwaser@digitalWisdomInstitute.org

Artificial intelligence, the "science and engineering of intelligent machines", still has yet to create even a simple "Advice Taker" [McCarthy 1959]. We have previously argued [Waser 2011] that this is because researchers are focused on problem-solving or the rigorous analysis of intelligence (or arguments about consciousness) rather than the creation of a "self" that can "learn" to be intelligent. Therefore, following expert advice on the nature of self [Llinas 2001; Hofstadter 2007; Damasio 2010], we embarked upon an effort to design and implement a self-understanding, self-improving loop as the totality of a (seed) AI. As part of that, we decided to follow up on Richard Dawkins' [1976] speculation that "perhaps consciousness arises when the brain's simulation of the world becomes so complete that it must include a model of itself" by defining a number of axioms and following them through to their logical conclusions. The results combined with an enactive approach yielded many surprising and useful implications for further understanding consciousness, self, and "free-will" that continue to pave the way towards the creation of safe/moral autopoiesis.

Keywords: autopoiesis; self; consciousness; ethics; free will; seed AI, safe AI, moral machines.

1. Introduction

John McCarthy coined the term artificial intelligence (AI) in creating a vision of breath-taking scope [McCarthy et al 1955]:

> *We propose that a 2 month, 10 man study of artificial intelligence be carried out [...] to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it [...] to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves.*

Yet, almost six decades later, it is not at all clear that we are anywhere close to being able to fulfill this grand vision – despite the fact that McCarthy's group originally believed that "*significant advance can be made […] if a carefully selected group of scientists work on it together for a summer.*" Clearly, the field is still overlooking something critical if we have no clear path towards success.

We have previously argued [Waser 2011] that the primary reason for the lack of progress is that the vast majority of AI researchers are far more focused on the analysis and creation of "intelligence" (problem solving and goal achievement)

rather than creating "a self" that can self-improve to intelligence. As Dennett [1994] points out

> *"Nobody doubts that any agent capable of interacting intelligently with a human being on human terms must have access to literally millions if not billions of logically independent items of world knowledge. Either these must be hand-coded individually by human programmers […] or some way must be found for the artificial agent to learn its world knowledge from (real) interactions with the (real) world.*

Recently some artificial general intelligence (AGI) researchers [Thorisson 2009] have answered this with what they call a "constructivist" (as opposed to "constructionist") approach by "*replacing top-down architectural design as a major development methodology with methods focusing on self-generated code and self-organizing architectures*". However, due to their replacement of the top-down approach, there still have still been no real efforts focused solely or primarily upon creating the top-level machine "selves" themselves – seemingly everyone expects them to self-organize.

The evolution of human consciousness leads us to expect that self-organization as well – but not before evolutionary time and population scales apply. Further, the self that initially appears may be far from safe or moral and there will likely be excessively dangerous tools produced as part of the process before then. As advocated by other machine ethics researchers [Wallach & Allen 2009], we believe that a top-down moral design must constrain the process for safety reasons. As Wallach [2010] states

> *Building moral machines is a practical, not a theoretical, goal. It is spurred by the need to ensure that increasingly autonomous machines will not cause harm to humans and other entities worthy of moral consideration.*

2. Self & Consciousness

In I Am a Strange Loop, Douglas Hofstadter [2007] argues that the key to understanding (our)selves is the "strange loop", a complex feedback network inhabiting our brains and, arguably, constituting our minds. Similar thoughts on self and consciousness are echoed by prominent neuroscientists [Llinas 2001; Damasio 1999, 2010]. Yet, despite all the smaller perception-action and cognitive cycles in AI and robotics, no has attempted to implement a self-understanding, self-improving loop as the totality of a (seed) AI embedded in a learning environment. Therefore, we [Waser 2012] decided to create such an entity using the LIDA Framework [Snaider et al 2011] based upon Baars' [1988] Global Workspace Theory of Consciousness as previously suggested [Wallach et al 2011] as being a good basis for conscious artificial moral agents.

As part of designing that effort, we decided to follow up on Richard Dawkins' [1976] speculation that "perhaps consciousness arises when the brain's simulation of the world becomes so complete that it must include a model of itself" by defining

Axiom 1. *The mind is a set of processes running on and fully instantiated in the physicality of the brain. This is exactly identical to the processes running on a "black-box" computer which the pathetic "fallacy"* ***correctly*** *labels as an entity.*

Axiom 2. *Consciousness is specifically those running processes of the mind that grow and alter the world-model and eventually automatize [Franklin et al] repetitive processes.*

Axiom 3. *Advanced consciousness/****selves*** *arise as Hofstadter's "strange loop" evolves/appears/self-creates.*

Axion 4. *Baars' [1997] Theater is simply our world model with our having been fooled into conflating our world model with the actual world and our self-model with our actual selves with Chalmers' [1995] double-aspect theory merely recognizing the presence of both the self and the created aspect of the self-model.*

Axiom 5. *The self-model is obviously the center of narrative gravity [Dennett 1992] and Dennett's [1991] "multiple drafts" are merely the inconsistencies inherent in trying to keep our representational model of ourselves up-to-date.*

Doing so revealed numerous surprising and useful implications for intelligence, self, consciousness, free will, and the possibility for a safe top-down moral design. For example, it has been asked [Torrance 2012] whether super-intelligence leads to super-consciousness. Given that AIXI [Hutter 2005] is theoretically maximally intelligent without being conscious (growing or altering its world model), the answer is clearly no. On the other hand, the greater the consciousness (ability to grow and perfect its world model), the more rapidly intelligence improves.

3. Grounding and Bounding

The biggest problems for and with artificial intelligence (AI) have always been those of grounding and bounding (or framing). Some fully specified (grounded and bounded) systems have had spectacular successes in endeavors ranging from beating chess grand masters to autonomous driving (of course, only to subsequently be declared not to be "true AI"). Other systems, thrown into environments more complex than expected (i.e. bounded but not completely specified), failed in equally spectacular fashion with attempts to ameliorate their "brittleness" turning into never-ending quagmires.

The term "Good Old-Fashioned AI (GOFAI)" was coined [Haugeland 1985] to declare that symbol manipulation alone was insufficient to create intelligence capable of dealing with the real world. The "frame problem" (which evolved from a formal AI problem [McCarthy & Hayes 1969] to a general philosophical question as

to how rational agents deal with the complexity and unbounded context of the world [Dennett 1984]), Searle's [1980] "Chinese Room", Harnad's [1990] "symbol grounding problem" and Dreyfus' [1972; 1979/1997; 1992] Heideggerian criticisms all seemingly prevent GOFAI from growing beyond closed and completely specified micro-worlds. We would argue that all of these problems are manifestations of a lack of either physical grounding and/or bounding or existential grounding and/or bounding.

Embodiment and physical grounding [Brooks 1990] initially seem to avoid or address some parts of these problems and have been enthusiastically embraced by many practitioners but it is clear that the larger portion of these problems remain. Physicality and relying upon the world itself to serve as its own model combine to automatically delineate a micro-world that ameliorates the symbol grounding and frame problems for purely physical tasks – but this merely finesses the current and local portion of the frame problem, rather than truly solving its totality. As a result, future and/or non-local tasks like learning and planning always turn out to be future concerns. Thus, even for purely physical tasks, Brooks [1997] himself notes that it certainly is the case that all biological systems are much more robust to changed circumstances and much quicker to learn or adapt than his systems – even going so far as to state "*The very term machine learning is unfortunately synonymous with a pernicious form of totally impractical but theoretically sound and elegant classes of algorithms*".

4. The Need for Intentionality

Further, Searle's "Chinese Room" was originally accompanied by a "robot reply" that he rejected as not making any substantial difference to his argument. As pointed by Haugeland [1981], our artifacts

> *only have meaning because we give it to them; their intentionality, like that of smoke signals and writing, is essentially borrowed, hence derivative. To put it bluntly: computers themselves don't mean anything by their tokens (any more than books do) - they only mean what we say they do. Genuine understanding, on the other hand, is intentional "in its own right" and not derivatively from something else.*

The problem with derived intentionality, as abundantly demonstrated by systems ranging from expert systems to robots, is that it is brittle and breaks badly as soon as it tries to grow beyond closed and completely specified micro-worlds and is confronted with the unexpected. As explained by Perlis [2010]

> *Brittle systems break when given a twist, when forced into circumstances beyond what they were explicitly designed for, i.e., when they encounter anomalies. This means our systems are far less useful than we would like, than we in fact need in order to deploy them usefully in realistic settings; for the real world abounds in the unexpected. And dealing usefully with the unexpected is a hallmark of flexible human-level intelligence, whereas*

> *breaking in the face of the unexpected (wasting time, getting nowhere, brooking disaster) is the hallmark of current automated systems. Rational anomaly-handling (RAH) is then the missing ingredient, the missing link between all our fancy idiot-savant software and human-level performance. Notice the boldness of this claim: not simply do our systems lack RAH, but this lack is the missing ingredient.*

We argue that rational anomaly handling is exactly and only being able to use and alter the system's world model. Or, in other words, it is consciousness as specified in axiom 2. Animals are only intentional/conscious to the extent that their very limited mental model can associatively tie together cause and effect (barking causes food or an open door) and that link, once learned, it is quickly automatized out of consciousness to a learned reflex. Meaning is provided solely by what "matters" to their physical body at the current time – as implemented by evolutionarily developed pleasant or unpleasant sensory arcs. True human-level planning intentionality requires that the world-model must somehow include its own meaning and genuine understanding so that it can sort out what "matters" and what should be pursued (grounding) and avoided (bounding).

Baum [2009] defines understanding a domain as "ability to rapidly produce computer programs to deal with new problems as they arise". While correct, we believe that this is too terse and will expand it as follows:

Definition 1: Understanding a domain is defined as the abilities

1. *to predict occurrences in the domain, most specifically including when and what you cannot predict, and*
2. *to rapidly produce plans/programs to deal with new problems as they arise in the domain.*

To do these two things, you obviously either need a complete set of domain-specific reflexes/programs (possible only in a closed, completely foreseen world) or a model of the domain to consult and build upon. If the domain includes yourself, you end up with Dawkins' consciousness.

5. Consciousness without Model or Self?

Uninformed debates over consciousness are a constant, huge thorn in the side of AI development – hemorrhaging time and sabotaging efforts. "Consciousness" is ***the*** primary example of what Marvin Minsky [2006] calls "suitcase words" – words that contain a variety of meanings packed into them. Theories of consciousness run from Panpsychism and Monism ("consciousness and the observed universe are exactly the same") to Dualism ("consciousness and the world are two completely different entities") and discussions about consciousness seemingly always devolve into debates about phenomenal consciousness, functional consciousness, qualia, and whether machines can even be conscious.

Even Daniel Dennett, who "wrote the book" on intentionality [1987] and consciousness [1991], made it very clear that he meant to avoid the debate as part of the MIT Cog project which aimed [1994],

> *not to make a conscious robot, but to make a robot that can interact with human beings in a robust and versatile manner in real time, take care of itself, and tell its designers things about itself that would otherwise be extremely difficult if not impossible to determine by examination.*

Cog was to "perform a lot of the feats that we have typically associated with consciousness in the past" without needing "to dwell on that issue from the outset." In order to reach that point, it was originally hoped that Cog would be "able to design itself in large measure, learning from infancy, and building its own representation of its world in the terms that it innately understands". For that reason, Cog's creators quite deliberately decided to "*make Cog as much as possible responsible for its own welfare*" and further, to equip it "with some innate but not at all arbitrary preferences, and hence provided of necessity with the concomitant capacity to be 'bothered' by the thwarting of those preferences, and 'pleased' by the furthering of the ends it was innately designed to seek."

What thwarted all of this is that Cog never really had anything that could really be called a "self". Indeed, it never even had any sort of internal world model. It had no sense of time and was never able to demonstrate coherent global (or unitary) behavior from the existing subsystems and sub-behaviors. The researchers spent a lot of time and effort on perceived pre-requisites but never found useful ways to even use much of its sensor information or solutions as to how episodic memory might arise. Yet, Dennett spoke as if Cog would innately have consciousness

> *Since we have cut off the dubious retreats to vitalism or origin chauvinism, it will be interesting to see if the skeptics have any good reasons for declaring Cog's pains and pleasures not to matter; at least to it, and for that very reason, to us as well. It will come as no surprise, I hope, that more than a few participants in the Cog project are already musing about what obligations they might come to have to Cog, over and above their obligations to the Cog team.*

Unfortunately, the Cog team never really worked out the implementation details of why its preferences mattered to Cog itself and why being 'bothered' or 'pleased' wasn't simply mere labeling and, again, derivative rather than innate. This reinforces Haugeland's [1985] claims that nothing could properly "matter" to an artificial intelligence and that mattering is crucial to consciousness and leads to Froese & Ziemke's [2009] question, "*how it is possible to design an artificial system in such a manner that relevant features of the world actually show up as significant from the perspective of that system itself, rather than only in the perspective of the human designer or observe?*"

We would argue that it should now be obvious that the first step is to create the self to have that perspective. Yet, Dennett (like most AI projects) seemingly persists

in the belief that parts will inevitably lead (magically self-assemble) to a whole and that an earwig is on a spectrum of intelligence that will eventually lead to humans. On the other hand, Brooks [1997] soberly concludes that

> *Perhaps there is a way of looking at biological systems that will illuminate an inherent necessity in some aspect of the interactions of their parts that is completely missing from our artificial systems . . . perhaps at this point we simply do not get it, and that there is some fundamental change necessary in our thinking in order that we might build artificial systems that have the levels of intelligence, emotional interactions, long term stability and autonomy, and general robustness that we might expect of biological systems.*

6. Evolution, Consciousness & Intentionality

As you go down the evolutionary scale, organisms and behavior get simpler and simpler. At the lowest level, behavior is strictly biochemical and/or mechanical and predictable with sufficient information (e.g. chemical gradients, etc.). Without sufficient information, however, environment-contingent behavior can appear to have exactly the same "as-if" intentionality as evolution itself. Unless the exact mechanisms can be perceived and understood, invoking the pathetic fallacy and saying that a bacteria "wants" the food that it is swimming towards or that evolution "wants" to create consciousness is a useful simplification for cognitive purposes. There is obviously some invisible mechanism or interiority that is causing identifiable self-correcting tendencies toward a particular result so we merely note the goal or target and move on. Unless, of course, we can perceive some regularities in the tendencies/behaviors that we can use to create even better predictions of how the subject will behave.

Evolution creates mindlessly through a process of generate and test. Better predictors and/or more aware entities are statistically more likely to survive and produce future generations. Thus, while the random mutations of evolution lack direction, this is certainly not true of evolution in general. With a few notable exceptions (like parasites), the preferential elimination of the less fit virtually always drives evolving systems towards increasing intelligence, complexity, integration, and capabilities. Aristotle's teleology of a seed was reproduction which we regard as "homopoiesis" (more resource intensive and certainly slower when implemented biologically – but also easier and safer than autopoiesis).

The existence of evolutionary "ratchets" (randomly acquired traits that are likely statistically irreversible once acquired due to their positive impact on fitness) causes "universals" of biological form and function to emerge, persist, and converge predictably even as the details of evolutionary path and species structure remain contingently, unpredictably different (Smart 2009). Ratchets can range from the broadly instrumental (enjoying sex) to the environmentally specific (streamlining and fins in water) to the contradictory and context-sensitive (like openness to

change). Skill at prediction extending over increasing lengths of time enabling planning and skill at prediction extending to one's self enabling self-improvement are what sets humans (and eventually intelligent machines) apart from animals in spite of our sub-par senses and awareness.

This meshes well with Tononi's [2004] information integration theory which argues that consciousness is one and the same thing as a system's capacity to integrate information (into the world model we would say). We would disagree, however, that the quantity of consciousness is the amount (in bits) of integrated information that flows through a given chokepoint rather than the amount of change applied to the world model. Tononi argues that the ability of a system to integrate information grows as that system incorporates statistical regularities from its environment and learns. Thus, when such information is about its environment, consciousness provides an adaptive advantage and evolved precisely because it is identical with the ability to integrate a lot of information in a short period of time.

Animals may be more aware of/reactive to their immediate surroundings (process more bits past a given point) but we humans are constantly aware of far more once the dimensions of time, distance/area, complexity and the recursive nature of self are all factored in – and much more able to predict and plan as a result. We have evolution to thank for that because evolution "wants" consciousness. Or is it that consciousness wanted to create itself out of the previously evolved simpler forms of biological self-creation and self-production of identity and meaning?

7. Enactivism

The enactive approach to cognitive science is generally considered to have coalesced with the publication of *The Embodied Mind* [Varela et al 1991], the introduction to which observed that the cognitive science of its time had "*virtually nothing to say about what it means to be human in everyday, lived situations*". As described [Torrance 2005a],

> *At the time when [it]was written, the primary focus of the interdisciplinary investigations associated with cognitive science was the nature of cognition, considered often in a rather narrow sense, as what humans do when they solve problems or seek to represent the world – the kinds of things that were relatively straightforward to model in (classical or connectionist) computer simulations. Since then the attention of the cognitive science community has broadened to include consciousness, emotion, dynamic embodied interaction with the world, and so on. In so doing it has come to be more closely in touch with every day, lived human experience.*

As described by Weber & Varela [2002], enactive cognitive science is essentially a synthesis of a long tradition of philosophical biology starting with Kant's "natural purposes" (or even Aristotle's teleology) and more recent developments in complex systems theory. Experience is central to the enactive approach and its primary distinction is the rejection of "automatic" systems, which

rely on fixed (derivative) exterior values, for systems which create their own identity and meaning. Critical to this is the concept of self-referential relations – the only condition under which the identity can be said to be intrinsically generated by a being for its own being (its self or itself).

Thompson [2004] describes Varela's path from cellular autopoiesis [Varela et al 1974] to biological autonomy [Varela 1979] to the continuity of life and mind [Maturana & Varela 1980, 1987] to *The Embodied Mind* [Varela et al 1991] to a biology of intentionality [Varela 1992] to the intertwining of identity and cognition [Varela 1997] to his final revised account of autopoiesis and autonomy [Weber & Varela 2002] which tied them in with Kant and the philosophy of biology of Hans Jonas [Jonas 1966/2001, 1968]. Thompson [2007] further describes the most salient of Varela's points as follows:

1. *Organizational closure refers to the self-referential (circular and recursive) network of relations that defines the system as unity*
2. *Operational closure refers to the reentrant and recurrent dynamics of such a system.*
3. *In an autonomous system, the constituent processes*
   (i) recursively depend on each other for their generation and their realization as a network,
   (ii) constitute the system as a unity in whatever domain they exist, and
   (iii) determine a domain of possible interactions with the environment.

Unfortunately, Varela limited the label of autopoiesis to the biochemical domain and the paradigmatic example of a single living cell and complained that others applying it elsewhere (most particularly to non-material systems like social institutions) "confuse autopoiesis with autonomy". We feel that this domain limitation is a grave error, first and foremost, because autopoiesis translated from the Greek literally has the meaning "self-creation" or "self-production" and it makes no sense to insist on further limiting that meaning without cause. Second, the term of art autonomy already has a very different definition (or set of definitions) in AI and robotics (more clearly specified as behavioral autonomy) and risks developing, if it hasn't already, the same plague of being a suitcase word that already bedevils consciousness and self. Finally, it leads to unhelpful over-specifications like the following [Torrance and Froese 2011]

> *What is it to be a (cognizing, conscious) agent? The five-fold response is as follows: it is (a) to be a biologically autonomous (autopoietic) organism – a precarious, far-from-equilibrium, self-maintaining dynamic system; (b) with a nervous system that works as an organizationally closed network, whose function is to generate significance or meaning, rather than (as in the "sense-model-plan-act" model) to act via a set of continually updated internal representations of the external world; (c) the agent's sense-making arises in virtue of the its dynamic sensorimotor coupling with its environment, such that (d) a world of significances is "enacted" or*

> *"brought forth" by a process whereby the enacted world and the organism mutually co-determine each other; and (e) the experiential awareness of that organism arises from its lived embodiment in the world.*

The obvious question being to ask why is necessary to be biological to be a "cognizing, conscious agent". Since it is much simpler to merely say that "consciousness is subsumptively and constitutionally autopoietic", we will continue, like others, to use that term contrary to Varela's wishes.

8. Identity & Morality

The most important contribution of enactivism for our purposes (of safety and morality) is the concept of self-constitution of identity ("constitutive autonomy") via "dynamic co-emergence" – an emergence not only through self-organization but also by self-production – where "the whole is constituted by the relations of the parts and the parts are constituted by the relations they bear to one another in the whole" [Thompson 2007]. It should be obvious that constitutive autonomy entails behavioral autonomy since it creates a self-correcting identity which is then the point of reference for the domain of interactions (i.e. "gives meaning"). Steve Torrance [2005b] says

Autopoiesis applies to self-maintaining agents of even the most primitive kind, yet it provides an essential element of what is involved in an adequate conception of highly developed, intelligent autonomous moral agency. Viewing beings as autonomous centres of meaning and purpose, as living and embodied conscious agents that enact their own existence, is, I believe, an important ingredient of building up a moral picture of ourselves, and those we wish to create in our moral image. On this picture, an agent will be seen as an appropriate source of moral agency only because of that agent's status as an self-enacting being that has its own intrinsic purposes, goals and interests. Such beings will be likely to be a source of intrinsic moral concern, as well as, perhaps, an agent endowed with inherent moral responsibilities. They are likely to enter into the web of expectations, obligations and rights that constitutes our social fabric. It is important to this conception of moral agency that MC agents, if they eventualize, will be our companions – participants with us in social existence – rather than just instruments or tools built for scientific exploration or for economic exploitability.

We intend that morality will be the primary part of the constitutive identity of the self that we will be creating (remembering as well that, from an autopoietic point of view, identity is analogous to, if not an actual restatement of, the relation necessary to bridge Hume's is-ought divide). Thus, we intend to create a self with an intention of being a moral entity that "suppresses or regulates selfishness and makes cooperative living possible" [Haidt 2010] and tries to help itself by helping the community build capability [Sen 1985, 2009; Nussbaum 2000, 2004].

It is also worth noting that "free will" is now easily defined as an entity's autopoietic identity (i.e. what it attempts to do and should succeed at doing outside

of extreme/unusual environmental influence). When unfortunate behavior arises, a useful question is whether it is easier to fix the environment or the autopoietic loop. In this view, punishment as deterrent is a fix to the environment while (attempting to) cure "insanity"/irrationality is (attempting to) fix the loop (and "temporary insanity" can be ignored since it most often is rational but normally anti-social behavior as a result of extreme, unlikely to repeat environment).

9. The Impossibility of Mary Dissolves the "Hard Problem"

Viewing consciousness through Dawkins' lens gives desperately needed insight on many existing problems – starting with then-dualist Frank Jackson's [1982] fiendishly simple puzzle for which we have yet to see the "obviously correct" analysis.

> *Mary is a brilliant scientist who is, for whatever reason, forced to investigate the world from a black and white room via a black and white television monitor. She specializes in the neurophysiology of vision and acquires, let us suppose, all the physical information there is to obtain about what goes on when we see ripe tomatoes, or the sky, and use terms like 'red', 'blue', and so on. ... What will happen when Mary is released from her black and white room or is given a color television monitor? Will she learn anything or not? It seems just obvious that she will learn something about the world and our visual experience of it. But then it is inescapable that her previous knowledge was incomplete. But she had all the physical information. Ergo there is more to have than that, and Physicalism is false.*

Using Dawkins' speculation, we would rephrase the puzzle as follows. Mary has a mental model of how her brain works. She acquires all the physical information there is to obtain about color perception and what happens as it is processed. Our question is, quite simply, "Where does she put all of this information?" How does her mind contain a model which is, necessarily, larger than it – since it not only contains the full specifications of the current mind with the current model but also all details of the behavior of the "black-box"/encapsulated sub-components of her mind? This is as bad as the recursive homunculus argument and seemingly, in the lively back and forth [Jackson 1986; Dennett 1991; Churchland 1990; Dennett 2006], no one has caught it (McGinn's [1989] "cognitive closure" was not directed at the correct target even though size problem invokes that concept).

The closest approach to the correct answer was provided by, appropriately enough, now-physicalist Frank Jackson [Garvey 2012]:

> *Looking red I think is clearly a representational state. I think the idea that perceptual states in general are representational states is extremely plausible. If you think that and you're a physicalist what you have to say is, right, Mary clearly enters a new representational state when she leaves the*

> *room. That should be common ground. If you're a physicalist, then you've got two things to say. You're either going to say, why doesn't she get new knowledge? Well, she already had it. If she already had it then you have to answer the question, what property do her newer experiences represent things as having which she knew about in the room? Maybe she didn't know about it under the name 'red', but if she's in a new representational state, and things are as they're being represented to be, and she doesn't learn anything new about the world, you need to give an answer to what looking red represents things as being, where the content of the representation can be expressed in physical terms. Alternatively, you can say it's a false representation. Colour is an illusion. You have to say one or the other.*

Inarguably, for human beings, color is an illusion. What we "see" has ***always*** been pre-processed by lower-level "black-box"/opaque processes. Since Mary cannot model/predict the representation that will occur, she ***will*** get new information because she literally cannot have all the physical information as the problem requires.

This obviously has great impact not only on what it is like to be a bat [Nagel 1974] and the constitutive autonomy that makes up its behavior but whether we can ever truly experience what it is like – since that experience must necessarily include **not** being capable of our own current mental capabilities as part of the experience. Note that this view also totally dissolves the "mystery" of the so-called "hard problem" of consciousness. We experience what we experience and we add a subset of our experience to our world model per the double aspect theory [Chalmers 1995]. There are two aspects to experience because we can't maintain a complete duplicate and we therefore necessarily create the second aspect.

In advance of experiencing something; however, we do not have the capacity to model the experience or how we will redact it to fit into our self-model. Qualia are "black-box" inputs to our system but they "feel" the way that they do because of the way in which they influence our state with Balduzzi & Tononi [2009] claiming, we believe correctly, that the network "shape" of the change matters but not claiming, as we believe, that the speed of the substrate matters critically. We can approximately describe the feeling after experiencing it by observing how it affects our self-model in comparison to other previous experiences – still remembering that we cannot predict it (like Mary).

10. The Possibility of Philosophical Zombies

Philosophical zombies are probably the best known problem in consciousness. According to Chalmers [1996], one can coherently conceive of an entire zombie world, a world physically indistinguishable from this world but entirely lacking conscious experience. The counterpart of every conscious being in our world would be a p-zombie. Since such a world is conceivable, Chalmers claims, it is logically possible, which is all, again he claims, that *his* argument requires. Chalmers also

states, though, that "*zombies are probably not naturally possible: they probably cannot exist in our world, with its laws of nature*."

We argue that Chalmers is wrong on almost every count. First, p-zombies regularly exist in our world. Every time you are "in the flow" or lose yourself in thought while driving, **you** are a p-zombie. In these cases, you may be ***externally*** physically indistinguishable; however, an fMRI will certainly show differences in what parts of the brain are active. If you aren't aware of fMRIs and their results, you exist in a world where your p-zombie state and your conscious state **are** physically indistinguishable ***to you*** and thus you have no problem conceiving of such a world – except that your conception is provably incorrect by teaching you about fMRIs and their results. Arguing that something is logically possible since ***you*** aren't aware of a counter-example is exactly akin to disproving a negative. Finally, a complete world of p-zombies would be completely brittle and completely collapse unless each p-zombie is maximally/AIXI intelligent.

11. Conclusion

As part of a project to create a safe/moral seed AI, we made a number of reasonable assumptions about consciousness to see where they led. Surprisingly, in combination with an enactive approach, they seemingly dissolved some of the most pernicious problems in consciousness and "free will" while apparently providing a self-correcting path towards a moral machine "self" that should be able to learn to, at least, a human level.

Acknowledgments

Thanks go to the Digital Wisdom Institute under whose auspices this article was created.

References

Balduzzi, B. & Tononi, G. [2009] "Qualia: The Geometry of Integrated Information," *PLoS Comput Biol* 5(8), e1000462. doi:10.1371/journal.pcbi.1000462.

Baum, E. B. [2009] "Project to build programs that understand" in *Artificial General Intelligence 2009: Proceedings of the Second AGI Conference (AGI '09)* (Arlington, VA) pp. 1-6.

Baars, B. [1997] *In the Theater of Consciousness: The Workspace of the Mind* (Oxford University Press).

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

Brooks, R. [1997] "From earwigs to humans," *Robotics and Autonomous Systems* **20**(2-4), pp. 291-304

Brooks, R. [1990] "Elephants don't play chess", *Robotics and Autonomous Systems* **6**(1-2), pp. 1-16.

Chalmers, D. [1996] *The Conscious Mind: In Search of a Fundamental Theory* (Oxford University Press).

Chalmers, D. [1995] "Facing Up to the Problem of Consciousness", *Journal of Consciousness Studies* **2**(3), pp. 200-219.

Churchland, P. [1989] *A Neurocomputational Perspective: The Nature of Mind and the Structure of Science* (MIT Press).

Damasio, A. R. [2010] *Self Comes to Mind: Constructing the Conscious Brain* (Pantheon).

Damasio, A. R. [1999] *The feeling of what happens: body and emotion in the making of consciousness* (Houghton Mifflin Harcourt).

Dawkins, R. [1976] *The Selfish Gene* (Oxford University Press).

Dennett, D. [2006] "What RoboMary Knows," in T. Alter & S. Walter (eds.), *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism* (Oxford University Press), pp. 15-31.

Dennett, D. [1994] "The practical requirements for making a conscious robot," *Phil Trans R Soc Lond A* **349**(1689), pp. 133–146.

Dennett, D. [1992] "The Self as a Center of Narrative Gravity," in F. Kessel, P. Cole & D. Johnson (eds) Self and Consciousness: Multiple Perspectives (Erlbaum). http://cogprints.org/266/.

Dennett, D. [1991] *Consciousness Explained* (Little Brown and Company).

Dennett, D. [1987] *The Intentional Stance* (MIT Press).

Dennett, D. [1984] "Cognitive Wheels: The Frame Problem of AI", in C. Hookway (ed.), *Minds, Machines, and Evolution: Philosophical Studies* (Cambridge University Press), pp. 129-151.

Dreyfus, H. L. [1992] *What Computers Still Can't Do: A Critique of Artificial Reason* (MIT Press).

Dreyfus, H. L. [1979/1997] "From Micro-Worlds to Knowledge Representation: AI at an Impasse," in J. Haugeland (ed.), *Mind Design II: Philosophy, Psychology, Artificial Intelligence* (MIT Press), pp. 143-182.

Dreyfus, H. L. [1972] *What Computers Can't Do: A Critique of Artificial Reason* (Harper & Row).

Franklin, S., Ramamurthy, U., D'Mello, S., McCauley, L., Negatu, A., Silva R., & Datla, V. [2007] "LIDA: A computational model of global workspace theory and developmental learning," In *AAAI Tech Rep FS-07-01: AI and Consciousness: Theoretical Foundations and Current Approaches (AAAI Fall Symposium '07)* (Arlington, VA), pp. 61-66.

Froese, T. & Ziemke, T. [2009] "Enactive Artificial Intelligence: Investigating the Systemic Organization of Life and Mind," *Artificial Intelligence*, **173**(3-4), pp. 466-500. doi: 10.1016/j.artint.2008.12.001.

Garvey, J. [2012] "Frank Jackson Interview," *The Philosophers' Magazine* **59**, pp. 66-75.

Haidt, J. & Kesebir, S. [2010] "Morality," in S. Fiske, D. Gilbert & G. Lindzey (Eds.) *Handbook of Social Psychology, 5th Edition* (2010), pp. 797-832.

Harnad, S. [1990] "The symbol grounding problem," *Physica D* **42**, pp. 335-346.

Haugeland, J. [1985] *Artificial Intelligence: The Very Idea* (MIT Press).

Haugeland, J. [1981] *Mind Design* (MIT Press/A Bradford Book).

Hofstadter, D. [2007] *I Am A Strange Loop* (Basic Books).

Hutter, M. [2005] *Universal Artificial Intelligence: Sequential Decisions Based on Algorithmic Probability* (Springer).

Jackson, F. [1982] "Epiphenomenal Qualia," *Philosophical Quarterly* **32**, pp. 127-36.

Jackson, F. [1986] "What Mary Didn't Know," *The Journal of Philosophy* **83**(5), pp. 291-295.

Jonas, H. [1966/2001] *The Phenomenon of Life: Toward a Philosophical Biology* (Northwestern University Press).

Jonas, H. [1968] "Biological Foundations of Individuality," *International Philosophical Quarterly* **8**, pp. 231-251.

Llinas, R. R. [2001] *I of the Vortex: From Neurons to Self* (MIT Press).

Maturana, H. R. & Varela, F. J. [1987] *The Tree of Knowledge: The Biological Roots of*

Human Understanding (Shambhala Publications).

Maturana, H. R. & Varela, F. J. [1980] *Autopoiesis and Cognition: The Realization of the Living* (Kluwer Academic Publishers).

McCarthy, J. [1959] "Programs with Common Sense," in *Mechanisation of thought processes* (NPL '58). http://www-formal.stanford.edu/jmc/mcc59/mcc59.html.

McCarthy, J. & Hayes, P. J. [1969] "Some philosophical problems from the standpoint of artificial intelligence," in B. Meltzer & D. Michie (eds.), *Machine Intelligence 4* (Edinburgh University Press), pp. 463-502.

McCarthy, J., Minsky, M., Rochester, N. & Shannon, C. [1955] A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence. http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html.

McGinn, C. (1989) "Can We Solve the Mind-Body Problem?", *Mind* **98**(391), pp. 349-366.

Minsky, M. [2006] *The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind* (Simon & Schuster).

Nagel, T. [1974] "What Is It Like To Be a Bat?", *The Philosophical Review* **83**, pp. 435-450.

Nussbaum, M. [2004] "Beyond the Social Contract: Capabilities and Global Justice," *Oxford Development Studies* **32(**1), pp. 3-18.

Nussbaum, M. [2000] *Women and Human Development: The Capabilities Approach* (Cambridge University Press).

Perlis, D. [2010] "BICA and Beyond: How Biology and Anomalies Together Contribute to Flexible Cognition," *International J. Machine Consciousness* **2**(2), pp. 1-11. doi: 10.1142/S1793843010000485.

Searle, J. [1980] "Minds, brains and programs," *Behavioral and Brain Sciences* **3**(3), pp. 417-457.

Sen, A. [2009] *The Idea of Justice* (Belknap/Harvard University Press).

Sen, A. [1985] *Commodities and Capabilities* (Oxford University Press).

Smart, J. M. [2009] "Evo Devo Universe? A Framework for Speculations on Cosmic Culture" in S. J. Dick & M. L. Lupisella (Eds.) *Cosmos and Culture: Cultural Evolution in a Cosmic Context, NASA SP-2009-4802* (USGPO), pp. 201-295.

Snaider, J., McCall, R.J. & Franklin, S. [2011] "The LIDA Framework as a General Tool for AGI," in *Artificial General Intelligence 2011: Proceedings of the Fourth AGI Conference (AGI '11)*(Mountain View, CA) 133-142

Thompson, E. [2007] *Mind in Life: Biology, Phenomenology, and the Sciences of Mind* (Belknap Press).

Thompson, E. [2004] "Life and Mind: From Autopoiesis to Neurophenomenology. A Tribute to Francisco Varela," *Phenomenology and the Cognitive Sciences* **3**, 381-398.

Thorisson, K. R. [2009] "From Constructionist to Constructivist A.I." in *AAAI Tech Report FS-09-01:Biologically-Inspired Cognitive Architectures (BICA '09)* (Arlington, VA) pp. 175-183.

Tononi, G. [2008] "Consciousness as Integrated Information: a Provisional Manifesto," *Biol. Bull.* **215**(3), pp. 216-242.

Tononi, G. [2004] "An Information Integration Theory of Consciousness," *BMC Neurosci.* **5**(42). doi:10.1186/1471-2202-5-42. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC543470/pdf/1471-2202-5-42.pdf.

Torrance, S. [2012] "Super-Intelligence and (Super-)Consciousness," *International J. Machine Consciousness* **4**(2), 1–19. doi: 10.1142/S179384301200098X.

Torrance, S. [2005a] "In search of the enactive: Introduction to Special Issue on Enactive Experience," *Phenomenology and the Cognitive Sciences* **4**(4), pp. 357-368.

Torrance, S. [2005b]. "Thin Phenomenality and Machine Consciousness," in *Proceedings of the Symposium on Next Generation Approaches to Machine Consciousness (AISB'05)*

(Hatfield, UK) pp. 59-66.

Torrance S. & Froese, T. [2011] "An Inter-Enactive Approach to Agency: Participatory Sense-Making, Dynamics, and Sociality," *Humana.Mente* **15**, pp. 21-54.

Varela, F. J. [1997] "Patterns of Life: Intertwining Identity and Cognition," *Brain and Cognition* **34**(1), pp. 72-87

Varela, F. J. [1992] "Autopoiesis and a Biology of Intentionality," in *Proc. of Autopoiesis and Perception: A Workshop with ESPRIT BRA 3352* (Dublin, Ireland), pp. 4-14

Varela, F. J. [1979] Principles of Biological Autonomy (Elsevier).

Varela, F. J., Maturana, H. R. & Uribe, R. [1974] "Autopoiesis: The organization of living systems, its characterization and a model", *BioSystems* **5**, pp. 187-196.

Varela, F. J., Thompson, E. & Rosch, E. [1991] The Embodied Mind: Cognitive Science and Human Experience (MIT Press).

Wallach, W. [2010] "Robot minds and human ethics: The need for a comprehensive model of moral decision making," *Ethics and Information Technology*, **12**(3), 243-250. doi: 10.1007/s10676-010-9232-8.

Wallach, W. & Allen, C. [2009]. *Moral machines: Teaching robots right from wrong* (Oxford University Press).

Wallach, W., Allen, C. & Franklin, S. [2011] "Consciousness and Ethics: Artificially Conscious Moral Agents," *International J. Machine Consciousness* **3**(1), 177–192. doi: 10.1142/S1793843011000674.

Waser, M. R. [2011] "Architectural Requirements & Implications of Consciousness, Self, and "Free Will"," in *Biologically Inspired Cognitive Architectures 2011: Proceedings of the Third Annual Meeting of the BICA Society (BICA '11)* (Arlington, VA) pp. 438-443. doi: 10.3233/978-1-60750-959-2-438.

Waser, M. R. [2012] "Safely Crowd-Sourcing Critical Mass for a Self-Improving Human-Level Learner/"Seed AI"," in *Biologically Inspired Cognitive Architectures 2012: Proceedings of the Third Annual Meeting of the BICA Society (BICA'12)* (Palermo, Sicily, Italy), pp. 345–350.

Weber, A. & Varela, F. J. [2002] "Life after Kant: Natural purposes and the autopoietic foundations of biological individuality," *Phenomenology and the Cognitive Sciences* **1**, pp. 97-125