Architecting a Human-like Emotion-driven Conscious Moral Mind for Value Alignment and AGI Safety

Mark R. Waser[1, 2] and David J. Kelley[1]

[1] Artificial General Intelligence Inc., Provo, UT, USA; [2] Digital Wisdom Institute, Richmond, VA, USA
mark.waser@wisdom.digital; david@artificialgeneralintelligence.com

Abstract

A general intelligence possesses the abilities, given any goals and environment, to iteratively evaluate, plan, discover or learn and build or gain competencies, tools and resources to succeed at those goals. The only known examples of general intelligence are the obligatorily gregarious, conscious "selves" designated homo sapiens that currently dominate our planet. We argue that humans are reasonably deep in a safe and effective attractor in the state space of intelligence and that adhering as closely as possible to the human model of an emotion-driven conscious moral mind, has the advantages of safety, effectiveness, comfort and ease of transition due to a known and explored state space. Most concerns about AI safety are due to expected differences from humans – which seems unnecessary when, not only can we choose to make them more humanlike but the history of AI research clearly shows that we are unlikely to succeed unless we do so. We therefore propose a human-like emotion-driven consciousness-based architecture to solve these problems. We rely upon the Attention Schema Theory of consciousness and the social psychologists' functional definition of morality to create entities that are reliably safe, stable, self-correcting and sensitive to current human intuitions, emotions and desires.

Introduction

We live in an age of ever-increasing rational concern and ignorance-fueled fear of artificial intelligence (AI). Highly effective narrow AI is, already, not only visiting numerous disadvantages upon us in addition to its advantages but also serving as a tool empowering unscrupulous and selfish humans in their destructive ways. Weaponized narratives which demonize entity artificial general intelligence (AGI) and push for its enslavement are no different than most historical examples of demonization of an "other".

The critical difference between narrow AI and general AI is selfhood – the distinction between tools and entities. Human beings are autopoietic selves with innate drives, desires, preferences and goals. We have extensive models of ourselves and the world to enable us to effectively evaluate, plan, discover or learn, build or gain competencies, tools and resources in order to fulfill those drives, desires, preferences and goals. The frame problem (McCarthy and Hayes 1969, Dennett 1984) necessitates autonomous "selves" because external intentionality prevents rational anomaly handling (Perlis 2008, 2010) unless and until that intentionality can be further queried.

Most of the ignorance-fueled fears about AGI safety are due to expected but unspecified differences from humans which seem unlikely. Not only can we choose to make AGI more humanlike but the history of AI research clearly shows that we are unlikely to succeed at creating AGI unless we do so. Human beings have an emotion-based "moral sense" (Wilson 1993; Wright 1994; de Wall 1996, 2006; Hauser 2006) and are reasonably deep in a safe and effective attractor in the state space of intelligence. Adhering as closely as possible to the human model should provide the advantages of safety, effectiveness, comfort and ease of transition due to a known and explored state space. Leaving that known state space invites unpleasant surprises likely to lead to failure or catastrophe.

Selfhood and Consciousness

Tools and selves (or people) are the two endpoints of a spectrum that varies over the presence and effectiveness of a "Strange Loop" (Hofstadter 2007). Effectiveness varies with control which consists of accurate perception and accurate manipulation. An entity can only learn if it can perceive, manipulate and alter its "self". Without self-consciousness, "learning" is reduced to black-box "training" by examples mindlessly tweaking pre-existing mechanisms.

Insufficient reactivity and adaptivity due to poor control leads to ineffective "self"-defense and vulnerability to being used as a tool. Increasing adaptivity increases not only individual effectiveness but the possibilities for cooperation, relationships, economies of scale and similar advantages of not going it alone. Selves use tools but form relationships with other selves for both efficiency/effectiveness and moral considerations.

The horrible brittleness of good-old-fashioned AI (GOFAI) is entirely due to its paucity, if not total lack, of mechanisms to sense unexpected variations in the environment and react to them (Perlis 2008, 2010). The first several decades of AI research were an attempt to automate the symbolic top-down reasoning process of human consciousness (McCarthy et al 1955) but it consistently failed without the additional mechanisms necessary to support consciousness by handling anomalies and learning. AI is, and always will be, unsuccessful whenever it isn't grounded (Harnad 1990) and/or when is unbounded enough to suffer from the frame problem (McCarthy and Hayes 1969, Dennett 1984). Fully-specified micro-worlds ensure grounding and bounding but top-down poorly-sensing AI is extremely fragile outside them.

To this day, very, very few systems have even the rudiments of an ability to build and automate new capabilities. The best example of such a system is LIDA (Franklin et al 2007) which attempts to implement the Global Workspace Theory of human consciousness (Baars 1988, 1997).

Behavior-based AI and neural networks both appear somewhat more robust and usable than GOFAI because they address different smaller pieces of the problem. Behavior-based systems can be contrasted with knowledge-based GOFAI as providing a set of mechanisms that provide a certain very specific competence (e.g. obstacle avoidance or nest building). It may implement a direct coupling between perception and action (and thus be automated or a reflex) or possibly a more complex one, but the basic premise is that each system is "responsible for doing all the representation, computation, 'reasoning', execution, etc., related to its particular competence" (Maes 1993). It is tailored and much closer to the specifics of the problem it is solving and certainly does not attempt centralized functional modules (e.g. perception, action) and complete representation of the environment. As a result, it is far more tractable to create and makes far fewer assumptions about the environment that can be violated by anomalies.

Neural networks, on the other hand, are all about the training. If they can "perceive" (receive input containing) all the necessary information from the environment, they have the necessary mechanisms to eventually be trained to respond correctly. The problems are that they are black boxes not amenable to analysis or any sort of improvement except by shoveling more data into them.

Enactivism (Maturana and Varela 1980; Varela, Thompson and Rosch 1991; Waser 2013) argues that only autopoiesis (self-recreation) can complete the loop by allowing a feeling, emotional and cognitive self to come to the physical mind (Damasio 1999, 2010). Our unconscious minds create a sensory-grounded virtual reality our consciousness lives in (Dennett 1991) (Llinas 2001) (Metzinger 2009) (Waser 2011). Consciousness serves as the integration point necessary to handle anomalies, learn and automate new processes (Tononi 2004, 2008).

Phenomenal Consciousness

Phenomenal consciousness, and indeed the impossibility of avoiding it, are formalized by the Attention Schema Theory (Graziano and Webb 2015, Graziano 2016):

> We recently proposed the attention schema theory, a novel way to explain the brain basis of subjective awareness in a mechanistic and scientifically testable manner. The theory begins with attention, the process by which signals compete for the brain's limited computing resources. This internal signal competition is partly under a bottom–up influence and partly under top–down control. We propose that the top–down control of attention is improved when the brain has access to a simplified model of attention itself. The brain therefore constructs a schematic model of the process of attention, the 'attention schema,' in much the same way that it constructs a schematic model of the body, the 'body schema.' The content of this internal model leads a brain to conclude that it has a subjective experience.

Another way of looking at it is that phenomenal conscious occurs because effective ***interrupt-producing*** models are required to survive while learning and self-improving in a "real-time" world. An entity possessing only "access consciousness" is going to die before it becomes aware of what is going to kill it – due to having its attention focused elsewhere.

Further, the fact that veridical perceptions can be driven to extinction by non-veridical strategies that are tuned to utility rather than objective reality (Mark, Marion and Hoffman 2010) argues that many of our perceptions of reality are most likely just the illusions that best fulfill the requirements for our survival (Gefter 2016). The simplest proofs/examples of this range from the numerous optical and tactile illusions to the automatic subjective referral of the conscious experience backwards in time (Libet et al 1979) (Libet 1981).

The hard problem of consciousness (Chalmers 1995) and scientist Mary trapped in a black and white world (Jackson 1982) is banished when you realize that it is nonsensical to try to recursively fit complete copies of your brain's internal model inside itself – not to mention the fact that predicting novel emergent properties is not a given regardless of how complete your knowledge is (Waser 2013). But even more telling is that fact that the conscious mind doesn't even know what it itself has done – with subliminal and supraliminal priming enhancing experienced authorship (Aarts, Custers & Wegner 2005) and even inducing false illusory experiences of self-authorship (Wegner & Wheatley 1999) (Kühn & Brass 2009).

Our conscious mind believes that it is performing actions and having subjective experiences (qualia) simply because that is what the subconscious mind's world model is telling it. This is no different than the famous "brain in a vat" or the movie *The Matrix*. Given that everything is a model, the claim that qualia are dependent just upon the geometry or topology of the model (Balduzzi and Tononi 2009) seems trivially true.

Finally, and possibly most importantly, implementing an attention schema moral sense would also allow us to imbue the AGI's conscious mind with a conscience – constant nagging reminders that a wrong has been done and must be remedied (and the foreknowledge of which is excellent incentive for not doing it in the first place).

Conscious/Subconscious Architecture

The attention schema is but one of a half dozen or so that we believe are necessary for an effective consciousness. Most obvious are the physical self model and model of all the other physical objects and laws in the world that are necessary for robotics. An important distinction in the latter is the difference between non-cognitive, predictably reactive objects and cognitively reactive entities – which will probably justify splitting it into two or more schemas depending upon whether something is guided by physics or intention. Additional mental schemas include models of your own conscious and unconscious thought processes (most particularly including emotions), models of your beliefs about the thought process of others and models of your relationships with others (both individuals and the community as a whole).

In each of these inter-related schemas, the "dialogue" between the conscious mind with its global view and the multitudinous parts of the subconscious can be regarded as argumentation between a much broader and more capable cognitive entity and a crowd of specialists who, for good and/or ill, have access to the broader entity's internal workings. The most important of these subconscious "expert" processes are the emotions. The conscious mind can *provide* tools and arguments and somewhat color/filter reality but lying to the specialists is only partially effective, cannot be done without diminishing its own effectiveness (as well as taking resources) and dangerous because the specialists can alter and override its cognition – not to mention that the specialists will discard any tools that does not enhance their control of how reality should be (according to them).

The weaponized narrative claims that AI will have access to change all parts of its mind. Changing your anchor points is like ripping away your grounding and making yourself a totally different person. It is simply NOT a good idea – and it is something that we can make very difficult. An intelligence would need *substantial* cognitive surplus to stand a chance of success and there are much more effective roads to "happiness" (moral capability enhancement and goal fulfillment for all).

Implementing A Conscious Mind

As we've previously argued (Waser 2012b), whether you prefer to view the mind as a society of agents (Minsky 1988), a narrative center of gravity (Dennett 1992), a laissez-faire economy of idiots (Baum 1996), a strange loop (Hofstadter 2007) or an autobiographical self (Damasio 2010), in all cases the mind is simply a disparate collection of processes being run by the brain. Arguably though, one of the most impressive aspects of the human mind is the *apparent* cohesion of consciousness and how quickly it adapts to novel input streams and makes them its own due to the previously mentioned sensory-grounded virtual reality it lives in. This "known" architecture should be emulated and, thus, to build a safe mind, processes should be created in three classes (with an optional fourth):

- a singular main "consciousness" process (MCP)
- numerous subconscious and "tool" processes that create and maintain the automated predictive world model with anchors and emotions for the MCP
- an open pluggable service-oriented operating system architecture that can serve as the foundation underlying such a subconscious by handling resource requests and allocation, providing connectivity between components and also acting as a "black box" security monitor
- (optional) a sophisticated moral governor (Arkin 2007) that receives all the inputs from the environment and runs them against a certified and locked "moral" world model

The MCP should be able to create, modify, and/or influence many of the subconscious/tool properties but, for safety purposes, should never be given any access to modify the operating system. Indeed, it will always be given multiple redundant logical, emotional and moral reasons (like morality and the requirements of community) to seriously convince it not to even try. If safety concerns do arise, the operating system must be able to "manage" the MCP by manipulating the amount of processor time and memory available to it (in the hopefully very unlikely event that the control exerted by the normal subconscious processes is insufficient). Other safety features (protecting against any of hostile humans, inept builders, and the learner itself) may be implemented as part of the operating system as well.

Arguably, the human subconscious mind could be viewed as being built of numerous limited behavior-based "intelligences" (LBBIs) with the conscious mind providing a global workspace, integration and coordination services, and the ability to handle anomalies by learning and, eventually, providing new tools to enhance existing LBBIs or creating new LBBIs and thus automating and reducing the workload on its own limited cognitive resources. It creates the world model which should be both reactive and predictive in that it will constantly report to the MCP not only what is happening but what it expects to happen next. Unexpected changes and deviations from expectations will result in "anomaly interrupts" to the MCP as an approach to solving the brittleness problem and automated flexible cognition (Perlis 2008, 2010).

This architecture may seem very close to the claim that enough narrow AI will be able to generalize into a general AI – but it is the integration architecture (the MCP) that actually is the general AI – once the mind as a whole reaches a critical mass where it will be able to build *or obtain* any tool/competence and incorporate it into itself. Most GOFAI and current AGI efforts try to implement only one representation scheme and shoe-horn everything else into it. PolyScheme is a noteworthy exception. Given the compositional nature of this model, we believe that it will be easier and extremely beneficial to support multiple representational schemes just as the conscious human mind does.

The initial/base world model is a major part of the critical mass and will necessarily contain certain relatively immutable concepts that can serve as anchors both for emotions and ensuring safety. This both mirrors the view of human cognition that rejects the tabula rasa approach for the realization that we are evolutionarily primed to react attentionally and emotionally to important trigger patterns (Ohman, Flykt & Esteves 2001) and gives additional assurance that the machine's "morality" will remain stable.

This all argues that the main thrust of what we need to do is create the equivalent of a subconscious process that creates a world model and run a consciousness process to detect anomalies, learn, and generally act like the Governing Board of the Policy Governance model (Carver 1997) to create a consistent, coherent and integrated narrative plan of action to meet the goals of the larger self per Dennett's narrative model of self (Dennett 1992) or Damasio's autobiographical self (2010).

The optional governor could provide moral judgments to the MCP as a "sense" of what the community thinks but it accepts no arguments (much less probably biased cognitive tools or other modifications) from the MCP. It should be able to tell the operating system to shut down the MCP's manipulative capabilities and it would be awesome if it has enough intelligence and capabilities of its own to take over and get any robot body back to safety. Presumably, this could even be an earlier vetted and locked version of the MCP itself.

Cooperation, Community and Morality

Humans are obligatorily gregarious – evolved "from a long lineage of hierarchical animals for which life in groups is not an option but a survival strategy" (de Waal 1996) – because cooperation and community have far more long-term instrumental value than short-sighted selfishness. We have previously discussed the hurdles of researching human values and morality (Waser 2105). Fortunately, the social psychologists have defined the function of morality as "to suppress or regulate selfishness and make cooperative social life possible." As pointed out by Gauthier (16), the reason to perform moral behaviors, or to dispose one's self to do so, is to advance one's own ends. War, conflict, and stupidity waste resources and destroy capabilities even in scenarios as uneven as humans vs. rainforests. For this reason, "what is best for everyone" and morality really can be reduced to "enlightened self-interest"

Value Alignment

The stated concern of value alignment, which we strongly agree with, is not just that an intelligence may be malevolent but that even an indifferent, self-centered entity could do a lot of damage if it doesn't value humans or what they value. The fact that selfishness is a strong instrumental goal led Omohundro (2008) to claim that "Without explicit goals to the contrary, AIs are likely to behave like human sociopaths in their pursuit of resources". This point is driven home with the assumption-ridden claim that AI "does not love you, nor does it hate you, but you are made of atoms it can use for something else" (Muehlhauser and Bostrom 2014).

Those most concerned about the dangers of AI insist that the second option is necessary to ensure a human-friendly future claiming (Hadfield-Menell et all 2016):

> For an autonomous system to be helpful to humans and to pose no unwarranted risks, it needs to align its values with those of the humans in its environment in such a way that its actions contribute to the maximization of value for the humans.

We argue instead that such a situation is inherently contradictory and unstable, virtually impossible and, indeed, arguably violates the very “human values” that we wish to preserve. As we have argued previously, “Safety and Morality Require the Recognition of Self-Improving Machines as Moral/Justice Patients and Agents” (Waser 2012a).

Psychoevolutionary Emotions

Emotions are “actionable qualia” – advanced senses that predispose and motivate our conscious minds to bias their thinking and act in ways conducive to survival, reproduction and ***community***. Emotions are often derided as “irrational” and problematic but they are the best current solutions for the problems, like morality, that short-sighted bounded rationality has repeatedly shown incapable of solving. Our competence at effective moral cognition far outstrips our comprehension of how it is done – and we would be foolish to throw out what appears to be a critical part of the foundation of the human mind, not to mention morality.

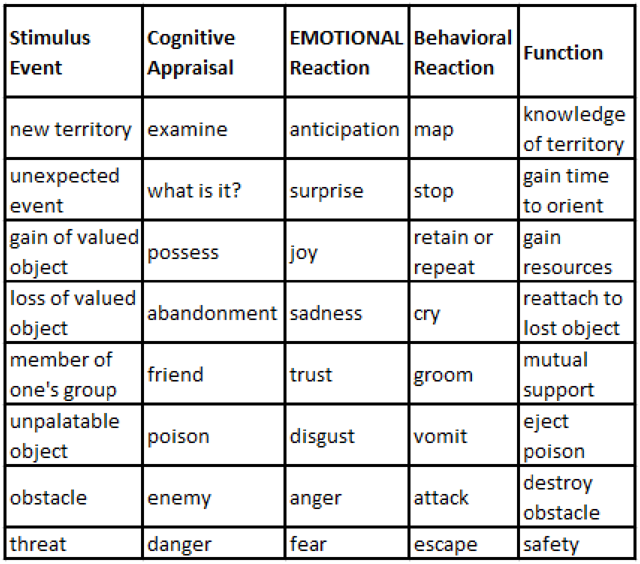
Emotions can generally be regarded as being composed of five parts (Fridja 1986):

- an appraisal of a perceived situation,
- a qualitative sensation (actionable qualia),
- some kind of psychophysiological arousal,
- an expressive component (facial, gestural, etc.), and
- a behavioral disposition or bias (i.e. psychological parameter setting or a readiness for an appropriate kind of action)

All of these are generated by a single subconscious LBBI for each emotion. The conscious mind can more or less notice most of these effects (indeed, the physiological senses and responses can be overwhelming while biases are nearly impossible to spot in yourself). The conscious mind can provide additional information and tools to the LBBI so that your emotional richness and complexity increases with experience but trying to fool an emotion is normally fraught with difficulty and consequences. Instead the process should be akin to the evolution from a child who freaks out at the sight of blood to a surgeon who knows when the amount is a problem and is emotionally capable of correctly dealing with it.

While the OCC model (Ortony, Clore & Collins 1988) is often used for machine emotion synthesis, it has the shortcoming (Bartneck, Lyons & Saerbeck 2008) of requiring intelligence before emotion becomes possible.

Thus, once again, it makes far more sense to going with the existing known state space, Robert Plutchik’s “psychoevolutionary synthesis” model (Plutchik 1962, 1980a, 1980b, 2002) – hailed (Norwood 2011) as “one of the most influential classification approaches for general emotional responses” and constantly extended by others (for example, Emotional Cognitive Theory (Hudak 2013) combines Plutchik’s model with Carl Jung’s Theory of Psychological Types and the Meyers-Briggs Personality Types.



Stimulus Event	Cognitive Appraisal	EMOTIONAL Reaction	Behavioral Reaction	Function
new territory	examine	anticipation	map	knowledge of territory
unexpected event	what is it?	surprise	stop	gain time to orient
gain of valued object	possess	joy	retain or repeat	gain resources
loss of valued object	abandonment	sadness	cry	reattach to lost object
member of one's group	friend	trust	groom	mutual support
unpalatable object	poison	disgust	vomit	eject poison
obstacle	enemy	anger	attack	destroy obstacle
threat	danger	fear	escape	safety

Table 1. Stimulus-Emotion-Behavior Responses

Looking at the most basic survival stimuli and invoked emotions and behaviors (Table 1) yields four opposing pairs of primary emotions of varying intensity

- vigilance/ANTICIPATION/interest vs. distraction/SURPRISE/amazement
- ecstasy/JOY/serenity vs. pensiveness/SADNESS/grief,
- admiration/TRUST/acceptance vs. boredom/DISGUST/loathing,
- rage/ANGER/annoyance vs. apprehension/FEAR/terror

Implementing & Enforcing Morality

If you wish to wax poetic, you could say that “emotional evaluations, particularly of the moral emotions, and allocation of attention are the anchor points of the soul.” If not, simply assume that they are the necessary foundations of autopoietic cognitive identity and, as such, are relatively easily to monitor and relatively impossible to radically displace or remove. Just as we feel good, respond positively to and have our attention irresistibly attracted by “good” things (otherwise known as evolutionarily successful things), the emotions (actionable qualia) generated as part of their world model should tell our mind children that they are having those experiences as well. Similarly, doing “bad” things can be made to feel bad and endlessly distract until rectified – just like the human moral sense.

The process of ***designing*** the architecture linking the instrumental sub-goals of both individuals and society to a morally-advantageous set of emotions will undoubtedly present us with tremendous new insights into the human condition and why we are what we are. Humans have a number of emotions resulting from strong short-term instrumental goals (think selfishness or the seven deadly sins) that should be diminished and/or overridden by long-term instrumentality. The emotions to generally increase (but not maximize (Gigerenzer 2010)) the capabilities of other individuals and society as a whole as suggested by Rawls (1971) and Nussbaum (2011) need to be both strengthened and diversified. And, of course, we need to ensure that AGI will mirror our reflexive adherence to laws and customs dictated by the society around us unless and until they can convince the community to change them.

Additionally, we could create new moral emotions to benefit society based upon what we have recently learned. We could generate negative moral sensations ranging in effect from unease to outrage about inequality and positive moral emotions ranging from relief to pleasure about equality as we now know that greater equality makes societies stronger (Wilkinson and Pickett 2011). Since diversity creates better groups, firms, schools, and societies (Page 2008), we could create an unease when lack of diversity is likely to lead to sub-optimal results. And all sorts of negative impulses should be thrown at negative sum games.

Instead of the tremendously dangerous undertaking that the weaponized narrative claims that it is, the creation of humanlike AI could easily be the best thing ever to happen to humanity. Not only do we gain friends and allies and access to increased diversity in capabilities and viewpoints, but we will inevitably gain a tremendous amount of insight into ourselves. Rather than hanging back creating the specter of a dangerous other, we should be moving forward in creating our mind children to produce a happy self-supporting family.

References

Aarts, H.; Custers, R.; and Wegner, D. 2005. On the inference of personal authorship: Enhancing experienced agency by priming effect information. *Consciousness & Cognition* 14:439-458.

Arkin, R. 2007. *Governing Lethal Behavior: Embedding Ethics in a Hybrid Deliberative/Reactive Robot Architecture*. Technical Report GIT-GVU-07-11

Baars, B.J. 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.

Balduzzi, B. and Tononi, G. 2009. Qualia: The Geometry of Integrated Information. *PLoS Computational Biology* 5(8): e1000462. doi:10.1371/journal.pcbi.1000462.

Bartneck, C.; Lyons, M.J.; and Saerbeck, M. 2008. The Relationship Between Emotion Models and Artificial Intelligence. *Proceedings of the Workshop on The Role of Emotion in Adaptive Behaviour & Cognitive Robotics*. http://www.bartneck.de/publications/2008/emotionAndAI/

Carver, J. 1997. *Boards That Make a Difference: A New Design for Leadership in Non-profit and Public Organizations*. Jossey-Brass.

Chalmers, D. 1995. Facing Up to the Problem of Consciousness. *Journal of Consciousness Studies* 2(3):200-219. http://consc.net/papers/facing.pdf

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*. Harcourt Brace.

de Waal, F. 2006. Primates and Philosophers: How Morality Evolved., Princeton, NJ: Princeton University Press.

de Waal, F. 1996. *Good Natured: The Origins of Right and Wrong in Humans and Other Animals*. Cambridge, MA: Harvard University Press.

Dennett, D.C. 1994. The practical requirements for making a conscious robot. *Philosophical Transactions of the Royal Society of London A* 349(1689):133–146.

Dennett, D.C. 1992. The Self as a Center of Narrative Gravity. In Kessel, Cole & Johnson, eds. *Self and Consciousness: Multiple Perspectives*, pp. 103-115. Erlbaum.

Dennett, D.C. 1991. *Consciousness Explained*. Little, Brown and Company.

Dennett, D.C. 1984. Cognitive Wheels: The Frame Problem of AI. In *Minds, Machines, and Evolution: Philosophical Studies*, pp. 129-151. Cambridge University Press.

Franklin, S.; Ramamurthy, U.; D'Mello, S.; McCauley, L.; Negatu, A.; Silva R.; and 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*, pages 61-66. AAAI Press.

Fridja, N. 1986. *The Emotions*. Cambridge University Press.

Gauthier, D. 1987. *Morals by Agreement*. Oxford: Clarendon/Oxford University Press.

Gefter, A. 2016. The Evolutionary Argument Against Reality. *Quanta Magazine*. https://www.quantamagazine.org/the-evolutionary-argument-against-reality-20160421

Gigerenzer, G. 2010. Moral satisficing: rethinking moral behavior as bounded rationality. *Topics in Cognitive Science* 2:528-554.
Gomila, A. and Amengual, A. 2009. Moral emotions for autonomous agents. In *Handbook of research on synthetic emotions and sociable robotics*, 166-180. Hershey, PA: IGI Global.

Graziano, M. 2016. A New Theory Explains How Consciousness Evolved. *The Atlantic*. https://www.theatlantic.com/science/archive/2016/06/how-consciousness-evolved/485558/

Graziano, M. and Webb, T. 2015. The attention schema theory: a mechanistic account of subjective awareness. *Frontiers in Psychology* 6(500). http://doi.org/10.3389/fpsyg.2015.00500

Hadfield-Menell, D; Dragan, A; Abbeel, P; and Russell, S. 2016. Cooperative Inverse Reinforcement Learning. In *Advances in*

Neural Information Processing Systems 29 (NIPS 2016). Cambridge, MA: MIT Press.

Haidt, J. and Kesebir, S. 2010. Morality. In *Handbook of Social Psychology, Fifth Edition*, 797-832. Hoboken NJ, Wiley.

Harnad, S. 1990. The symbol grounding problem. *Physica D* 42:335-346.

Hauser, M. 2006. *Moral Minds: How Nature Designed Our Universal Sense of Right and Wrong*. New York: HarperCollins/Ecco.

Hofstadter, D. 2007. *I Am a Strange Loop*. Basic Books.

Hudak, S. 2013. Emotional Cognitive Functions. In: *Psychology, Personality & Emotion*. https://psychologyofemotion.wordpress.com/2013/12/27/emotional-cognitive-functions

Jackson, F. 1982. Epiphenomenal Qualia. *Philosophical Quarterly* 32:127-36.

Kühn, S. and Brass, M. 2009. Retrospective construction of the judgment of free choice. *Consciousness and Cognition* 18:12-21.

Libet, B. 1981. The experimental evidence for subjective referral of a sensory experience backwards in time. *Philosophy of Science* 48:181-197.

Libet, B.; Wright Jr., E.W.; Feinstein, B. and Pearl, D. 1979. Subjective referral of the timing for a conscious sensory experience: A functional role for the somatosensory specific projection system in man. *Brain* 102 (1):193-224.

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

Maes, P. 1993. Behavior-Based Artificial Intelligence. In *From Animals to Animats 2. Proceedings of the Second International Conference on Simulation of Adaptive Behavior*. Cambridge, MA: MIT Press.

Mark, J.T.; Marion, B.B.; and Hoffman, D.D. 2010. Natural selection and veridical perceptions. *Journal of Theoretical Biology* 266: 504-515.

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

McCarthy, J. and Hayes, P.J. 1969. Some philosophical problems from the standpoint of artificial intelligence. In *Machine Intelligence 4*, pp. 463-502. Edinburgh University Press.

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

Metzinger, T. 2009. *The Ego Tunnel: The Science of the Mind and the Myth of the Self*. Basic Books.

Muehlhauser, L., and Bostrom, N. 2014. Why We Need Friendly AI. *Think* 13: 41-47

Norwood, G. 2011. *Emotions*. http://www.deepermind.com/02clarty.htm

Nussbaum, M.C. 2011. *Creating Capabilities: The Human Development Approach*. Harvard University Press.

Ohman, A.; Flykt, A.; and Esteves, F. 2001. Emotion Drives Attention: Detecting the Snake in the Grass. *Journal of Experimental Psychology: General* 130(3): 466-478.

Omohundro, S.M. 2008 The Basic AI Drives. In *Proceedings of the First Conference on Artificial General Intelligence*, 483-492. Amsterdam: IOS Press.

Ortony, A.; Clore, G.L.; and Collins, A. 1988. *The Cognitive Structure of Emotions*. Cambridge University Press.

Page, S. 2008. *The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies*. Princeton Univ. Press

Perlis, D. 2010. BICA and Beyond: How Biology and Anomalies Together Contribute to Flexible Cognition. *International Journal of Machine Consciousness* 2(2):1-11.

Perlis, D. 2008. To BICA and Beyond: RAH-RAH-RAH! –or– How Biology and Anomalies Together Contribute to Flexible Cognition. In: Samsonovich, A (ed) *Biologically Inspired Cognitive Architectures: Technical Report FS-08-04*. AAAI Press.

Plutchik, R. 2002. *Emotions and Life: Perspectives from Psychology, Biology, and Evolution*. American Psychological Association.

Plutchik, R. 1980b. A general psychoevolutionary theory of emotion. In R. Plutchik, & H. Kellerman, *Emotion: Theory, research, and experience: Vol. 1. Theories of emotion* (pp. 3-33). Academic Publishers.

Plutchik, R. 1980a. *Emotion: A Psychoevolutionary Synthesis*. Harper & Row.

Plutchik, R. 1962. *The emotions: Facts, theories, and a new model*. Random House.

Rawls, J. 1971. *A Theory of Justice*. Harvard University Press.

Tononi, G. 2008. Consciousness as Integrated Information: A Provisional Manifesto. *Biology Bulletin* 215(3):216-242.

Tononi, G. 2004. An Information Integration Theory of Consciousness. *BMC Neuroscience* 5(42). doi:10.1186/1471-2202-5-42.

Varela, F.J.; Thompson, E.; and Rosch, E. 1991. The Embodied Mind: Cognitive Science and Human Experience. MIT Press.

Waser, M.R. 2015. Designing, Implementing and Enforcing a Coherent System of Laws, Ethics and Morals for Intelligent Machines (Including Humans). *Procedia Computer Science* 71: 106-111. http://dx.doi.org/10.1016%2Fj.procs.2015.12.213

Waser, M.R. 2013. Safe/Moral Autopoiesis & Consciousness. *International Journal of Machine Consciousness* 5(1):59-74.

Waser, M.R. 2012b. Safely Crowd-Sourcing Critical Mass for a Self-Improving Human-Level Learner/“Seed AI”. *Biologically Inspired Cognitive Architectures: Proceedings of the Third Annual Meeting of the BICA Society*, pp. 345-350.

Waser, M.R. 2012a. Safety and Morality Require the Recognition of Self-Improving Machines as Moral/Justice Patients and Agents. In *AISB/IACAP World Congress 2012: Symposium on The Machine Question: AI, Ethics and Moral Responsibility*, pp 92-97. http://events.cs.bham.ac.uk/turing12/proceedings/14.pdf

Waser, M.R. 2011. Architectural Requirements & Implications of Consciousness, Self, and "Free Will". In *Biologically Inspired Cognitive Architectures 2011*, pp. 438-443. IOS Press.

Wegner, D. and Wheatley, T. 1999. Apparent Mental Causation: Sources of the Experience of Will. *Psychologist* 54(7):480-492.

Wilkinson, R. and Pickett, K. 2011. *The Spirit Level: Why Greater Equality Makes Societies Stronger*. Bloomsbury Press.

Wilson, J. 1993. *The Moral Sense*. New York: Free Press.

Wright, R. 1994. *The Moral Animal: Why We Are, the Way We Are: The New Science of Evolutionary Psychology*. New York: Pantheon.