Architectural Requirements & Implications of Consciousness, Self, and "Free Will"

Mark WASER
Books International

Abstract. While adaptive systems are currently generally judged by their degree of intelligence (in terms of their ability to discover how to achieve goals), the critical measurement for the future will be where they fall on the spectrum of self. Once machines and software are able to strongly modify and improve themselves, the concepts of the self and agency will be far more important, determining not only what a particular system will eventually be capable of but how it will actually act. Unfortunately, so little attention has been paid to this fact that most people still expect that the basic cognitive architecture of a passive "Oracle" in terms of consciousness, self, and "free will" will be little different from that of an active explorer/experimenter with assigned goals to accomplish. We will outline the various assumptions and trade-offs inherent in each of these concepts and the expected characteristics of each – which not only apply to machine intelligence but humans and collective entities like governments and corporations as well.

Keywords. Intelligence, Consciousness, Self, Agency, "Free Will", Morality

Introduction

Imagine three futures. In one, machine intelligence develops gradually as computers continue growing more varied and more ubiquitous leading to a tremendous variety of sentient entities. In another, a single machine intelligence suddenly appears, quickly spreads to every interconnected computer, and eventually controls literally billions of androids and other machines. In the third, mankind creates a nearly omniscient machine "oracle" that gives humanity tremendous power and control over their lives.

In determining the future of humanity, the frequently overlooked concept of self is likely to be one of the most critical factors. Indeed, the coalescence and increasing intelligence of larger than human entities has already had a tremendous impact on humanity which is only accelerating. Tribes combining into city-states merging into countries vying with international corporations shape the world that we live in. Political parties, media conglomerates, grassroots movements and "larger-than-life" individuals (real and fictional) all fight to shape our thoughts. The increasing size and speed of flash mobs and the intelligence displayed by the self-named "HiveMind" created by "I Love Bees" [1] are undoubtedly only portents of things to come.

Yet, the vast majority of AGI researchers are far more focused on the analysis and creation of intelligence rather than self and pay little heed to the differences between a passive "oracle", which is frequently perceived as not possessing a "self", and an active autonomous explorer, experimenter, and inventor with specific goals to accomplish. Arguably, however, it is the "self" that was co-created with biological intelligence – and it is the goals and motivations of any "self" that exists that will determine the

behavior of future machine intelligences. Current predictions of the future vary from tacitly expecting individual selves in androids to believing that a single Borg- or Skynet-like hive self is unavoidable to insisting that the non-existence of self, in terms of independent motivation and free will, is absolutely required for human safety.

Compounding the issue is there is no still concrete consensus on what self and/or conscious are or how they arise. If we wish to insist upon the non-existence of self, what are the lines that we should not cross and the things that we should not build? Is self-modification possible without a true self? Does allowing "self"-modification usually or inevitably lead to a self? An examination of these issues is long overdue.

1. Consciousness

AGI researchers seem to have converged on a definition of intelligence as a measure of the ability to determine how to achieve a wide variety of goals under a wide variety of circumstances. Or, alternatively, that the function of intelligence is to determine the method(s) by which a wide variety of goals can be achieved under a wide variety of circumstances. The degree of intelligence, therefore, is the minimal amount of information processing necessary to determine how to manipulate the circumstances so they include the goal.

Chalmer's ***double-aspect theory*** of information [2] claims that the fact that "there is a direct isomorphism between certain physically embodied information spaces and certain phenomenal (or experiential) information spaces" (or, alternatively, that "we can find the same abstract information space embedded in physical processing and in conscious experience") means that the experience of consciousness is created by the structure of information processing. This leads to statements and speculation that

> Where there is simple information processing, there is simple experience, and where there is complex information processing, there is complex experience. A mouse has a simpler information-processing structure than a human, and has correspondingly simpler experience; perhaps a thermostat, a maximally simple information processing structure, might have maximally simple experience?

This meshes well with Tononi's ***information integration theory*** of consciousness which argues that subjective experience is one and the same thing as a system's capacity to integrate information, that the quantity of consciousness is the amount of integrated information generated by a complex of elements [3] and that the quality of experience (qualia) is specified by (the geometry of) the informational relationships it generates [4]. Tononi argues that the ability of a system to integrate information grows as that system incorporates statistical regularities from its environment and learns. Thus, if such information is about its environment, consciousness provides an adaptive advantage and may have evolved precisely because it is identical with the ability to integrate a lot of information in a short period of time. And since intelligence is the ability to integrate and manipulate information, consciousness seems a pre-requisite for intelligence and unavoidable in AI.

2. Self and Sense of Self

If consciousness is a foregone conclusion, the next best and/or only way to ensure AGI safety is to ensure that they either don't have a self or don't have knowledge of

their own self. For example, a passive Oracle that does nothing except answer questions is generally not considered a danger. Except that a "passive" Oracle either needs to collect and integrate information in order to be able to answer questions or needs a side-kick that does so. And the process(es) that gather and integrate the information will undoubtedly have goals that should include timeliness, accuracy and safety. If the system is conscious/aware of these goals (i.e. they are integrated in with the rest of the information available to the system), for questions that are large and/or long-term enough, it will undoubtedly be most effective for the process(es) to first optimize or improve itself, if the system is able to do so, and then collect the information and answer the question (or intersperse and alternate the various activities).

This complete loop of a process (or physical entity) modifying itself must, if indeterminate in behavior, necessarily and sufficiently be considered an entity rather than an object – and humans innately tend to do so with the pathetic fallacy. In "I Am a Strange Loop", Hofstadter [5] talks about self, soul, consciousness, and the concept of "I" as if they are the same thing. Baars [6] writes of self as "unifying context of consciousness", "overall, unifying context of personal experience", and "a framework that remains largely stable across many different life situations" which "like any context, self seems to be largely unconscious but it profoundly shapes our conscious thoughts and experiences." He writes of consciousness as "gateway to the unconscious mind" and as "gateway to the self" with the function "to create access for the self in all its manifestations" and quotes Daniel Dennett [7] calling it "that to which I have access." Dennett also [8] describes the self as "a center of narrative gravity".

The potential problems with self-modifying goal-seeking entities are amply described by Omohundro [9]. In addition to the self-improvement described above, such an entity will have other tendencies (instrumental goals) which Omohundro claims will be present unless they are explicitly counteracted because they, too, advance any other goals present. These include rationality, effective evaluation, avoiding manipulation, self-protection, and acquisition and effective use of resources. Unfortunately, because he failed to recognize co-operation and morality as instrumental goals, Omohundro claimed that "Without explicit goals to the contrary, AIs are likely to behave like human sociopaths in their pursuit of resources." We would argue that an entity with enough goals and intelligence will eventually reach such recognition but likely that it could become powerful enough to destroy us before that point.

3. Autonomy, Responsibility and "Free Will"

The critical safety issues that really must be addressed are responsibility for safety, how safety will be ensured, and how the responsibility will be enforced. Defining autonomy as "freedom to determine one's own actions, behaviour, etc." and responsibility as "the ability or authority to act or decide on one's own, without supervision" implies an agent that "determines"/"decides" without outside interference.

Some people have suggested that all that is necessary for safety is to insert a human into the loop of self-modification (i.e. that a knowledge/sense of "self" is safe as long as the self-loop isn't fully integrated). This places the onus of responsibility on the human and assumes that they will take the time to understand the system and any proposed modifications well enough not to allow any dangerous modifications to pass (and not be out-classed enough in intelligence that the system could "dupe" them into doing so). At best, this will "only" tremendously slow the rate of system improvement.

At worst, it will be entirely ineffective. A far better strategy would be to create a safety evaluation or morality process/self whose entire goal/purpose/self is solely to identify and point out dangers in proposed answers (including modifications).

Eliezer Yudkowsky's Friendly AI [10] makes the programmer responsible for safety by insisting upon an absolutely foolproof "Friendliness Structure" that will ***forever*** convince the autonomous AI that it "wants" to be safe because of the goal that was initially planted in it. In his model, the AI will quickly become too powerful for anything to be externally enforced on it and post hoc measures against either it or the programmer are pretty much pointless. Yudkowsky later [11] took to calling his Friendly AI a "Really Powerful Optimization Process", presumably to avoid questions of entity-ness, selfhood, free will and slavery.

Alan Felthous [12] says that "Free will is regarded by some as the most and by others as the least relevant concept for criminal responsibility." Anthony Cashmore [13] claims that "a basic tenet of the judicial system and the way that we govern society is that we hold individuals accountable (we consider them at fault) on the assumption that people can make choices that do not simply reflect a summation of their genetic and environmental history." He quotes de Duve [14] arguing that if "neuronal events in the brain determine behavior, irrespective of whether they are conscious or unconscious, it is hard to find room for free will. But if free will does not exist, there can be no responsibility, and the structure of human societies must be revised." Yet, as Felthous points out, leading historical jurists in England eventually dropped the descriptor "free" but retained the central importance of the will to criminal responsibility and emphasized its dependence on the intellect to function properly.

Today, most of the information about will and morality is coming from the neurosciences, indicating that much of what we believe we directly experience and will is actually generated unconsciously and/or revised post hoc. For example, consciousness always edits out the approximately one-half second time delay between when physical stimulus first appears in the appropriate sensory region of the brain and when it actually enters conscious awareness with experiments [15][16] clearly demonstrating that there is an automatic subjective referral of the conscious experience backwards in time. More interestingly, not only have numerous studies [17][18][19] shown that the cerebral activity for action starts well before conscious intention but revealed that the upcoming outcome of a decision could be found in study of the brain activity in the prefrontal and parietal cortex up to 7 seconds before the subject was aware of their decision [20] and that the perceived time of decision is inferred rather than sensed and can be altered by deceptive feedback [21] or belief in personal or other human agency as opposed to that of machines [22][23].

More surprisingly, studies show that even agency is inferred rather than sensed with subliminal and supraliminal priming enhancing experienced authorship [24] and even inducing false illusory experiences of self-authorship [25][26] with belief in "free will" being enhanced by both [27]. This is obviously useful since psychological studies repeatedly proved that low control belief affects performance and motivation and recent studies have even shown that undermining beliefs in free will affects brain correlates of voluntary motor preparation by reducing action potential more than a second before subjects consciously decided to move [28] and increased cheating [29].

For longer-term activities, there is ample evidence [30] to show that our conscious, logical mind is constantly self-deceived to enable us to most effectively pursue what appears to be in our own self-interest and other recent evidence [31] clearly refutes the common assumptions that moral judgments are products of, based upon, or even

correctly retrievable by conscious reasoning. We don't consciously know and can't consciously retrieve why we believe what we believe and are actually even very likely to consciously discard the very reasons (such as the "contact principle") that govern our behavior when unanalyzed. Of course, none of this should be particularly surprising since Minsky [32] has pointed out many other examples, as when one falls in love, where the subconscious/emotional systems overrule or dramatically alter the normal results of conscious processing without the consciousness being aware of the fact. Yet, arguably, we have evolved these features because they work well and make us more evolutionarily "fit". Maybe the best design for a safety evaluation and/or morality process is the one that we humans model – and maybe we should attempt it first for our machine, our corporations, and our governments (or, at least, understand it).

Conclusion – Safety in Stories and Illusions

Susan Blackmore [33] describes consciousness as an illusion of "a continuous stream of rich and detailed experiences, happening one after the other to a conscious person". Whenever consciousness is required, "a retrospective story is concocted about what was in the stream of consciousness a moment before, together with a self who was apparently experiencing it. Of course there was neither a conscious self nor a stream, but it now seems as though there was. This process goes on all the time with new stories being concocted whenever required.

This matches well with Dennett's [34] theory of multiple drafts that at any time there are multiple constructions of various sorts going on in the brain - multiple parallel descriptions of what's going on. None of these is 'in' consciousness while others are 'out' of it. Rather, whenever a probe is put in - for example a question asked or a behaviour precipitated - a narrative is created. The rest of the time there are lots of contenders in various stages of revision in different parts of the brain, and no final version. As he puts it "there are no fixed facts about the stream of consciousness independent of particular probes".

So apparently, safety lies in stories and illusions. For humans, Albert Bandura [35] has recommended reframing the issue of free will in terms of the exercise of agency, operating principally through cognitive and other self-regulatory process to provide new insights into the constructive and proactive role that cognition plays in action. We would argue that his social cognitive theory which embeds intelligences in a society is well-grounded, well-explored, adaptable to all intelligences, and a useful perspective for starting to understand everything from humans to machines to corporations and governments.

References

[1] J. McGonigal, Why I Love Bees: A Case Study in Collective Intelligence Gaming, in K. Salen (ed.), *The Ecology of Games: Connecting Youth, Games, and Learning*, 199-228, The MIT Press, Cambridge, MA, 2008, doi: 10.1162/dmal.9780262693646.199.

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

[3] G. Tononi, An Information Integration Theory of Consciousness. *BMC Neurosci*. **5:42** (2004)

[4] B. Balduzzi and G. Tononi, Qualia: The Geometry of Integrated Information. *PLoS Comput Biol* **5:8** (2009), e1000462. doi:10.1371/journal.pcbi.100046

[5] D. Hofstadter, *I Am A Strange Loop*, Basic Books, New York, NY, 2007.
[6] B. Baars, *In The Theater Of Consciousness: The Workspace of the Mind*, Oxford University Press, New York, NY, 1997.
[7] D. Dennett, *Consciousness Explained*, Little, Brown, and Co., New York, NY, 1991.
[8] D. Dennett, The Self as a Center of Narrative Gravity, In F. Kessel, P. Cole & D. Johnson (eds) *Self and Consciousness: Multiple Perspectives*, Erlbaum, Hillsdale, NJ, 1992, http://cogprints.org/266/.
[9] S. Omohundro, The Basic AI Drives, In p. Wang, B. Goertzel, S. Franklin (eds) *Proceedings of the First Conference on Artificial General Intelligence*, 483-492. Amsterdam: IOS Press.
[10] E. Yudkowsky, *Coherent Extrapolated* Volition. 2004. http://www.singinst.org/upload/CEV.html.
[11] E. Yudkowsky, *Creating Friendly AI 1.0: The Analysis and Design of Benevolent Goal Architectures*, http://singinst.org/CFAI.html.
[12] Alan R. Felthous, MD, The Will: From Metaphysical Freedom to Normative Functionalism, *J Am Acad Psychiatry Law* **36:1** (2008), 16-24
[13] A. Cashmore, The Lucretian swerve: the biological basis of human behavior and the criminal justice system, *Proc Natl Acad Sci U S A* **107:10** (2010), 4499-504
[14] C de Duve, *Vital Dust: The Origin and Evolution of Life on Earth*, Basic Books, New York, NY, 1995.
[15] S. Blackmore, There is no stream of consciousness, *Journal of Consciousness Studies* **9:5** (2002), 17-28
[16] D. Dennett, Toward A Cognitive Theory of Consciousness, in D. Dennett (ed) *Brainstorms*, 149-173, Bradford Books/MIT Press, Cambridge, MA, 1978
[17] Libet, B., Wright, E. W., Feinstein, B., and Pearl, D. Subjective referral of the timing for a conscious sensory experience: A functional role for the somatosensory specific projection system in man, *Brain* **102:1** (1979), 193-224
[18] B. Libet, The experimental evidence for subjective referral of a sensory experience backwards in time: Reply to P. S. Churchland. *Philosophy of Science* **48** (1981), 181-197.
[19] B. Libet, C. A. Gleason, E. W. Wright, and D. K. Pearl, Time of conscious intention to act in relation to onset of cerebral activity (readiness-potential): The unconscious initiation of a freely voluntary act. *Brain* **106** (1983), 623-642.
[20] B. Libet, Unconscious cerebral initiative and the role of conscious will in voluntary action, *Behavioral and Brain Sciences* **8** (1985), 529-539 DOI: 10.1017/S0140525X00044903
[21] M. Matsuhashi and M. Hallett, The timing of the conscious intention to move. *Eur J Neurosci*. **28:11** (2008), 2344-51.
[22] C. Soon, M. Brass, H. Heinze and J. Haynes, Unconscious determinants of free decisions in the human brain, *Nature neuroscience* **11:5** (2008), 543–545. doi:10.1038/nn.2112
[23] Banks and E. Isham, We Infer Rather Than Perceive the Moment We Decided to Act, *Psychological Science* **20:1** (2009), 17-21
[24] A. Wohlschläger, P. Haggard, B. Gesierich and W. Prinz1, The Perceived Onset Time of Self- and Other-Generated Actions, *Psychological Science* 14:6 (2003) 586-591
[25] M. Buehner and Gruffydd R. Humphreys, Causal Binding of Actions to Their Effects, *Psychological Science* **20:10** (2009), 1221-1228, doi: 10.1111/j.1467-9280.2009.02435.x
[26] H. Aarts, R. Custers and D. Wegner, On the inference of personal authorship: Enhancing experienced agency by priming effect information. *Consciousness & Cognition* **14** (2005), 439-458
[27] D. Wegner and T. Wheatley, Apparent Mental Causation: Sources of the Experience of Will. *American Psychologist* **54:7** (1999), 480-492.
[28] S. Kühn and M. Brass, Retrospective construction of the judgment of free choice. *Consciousness and Cognition* **18** (2009), 12-21.
[29] H. Aarts and K. van den Bos, On the Foundations of Beliefs in Free Will: Intentional Binding and Unconscious Priming in Self-Agency, *Psychological Science* **22:4** (2011), 532-537
[30] D. Rigoni, S. Kühn, G. Sartori and M. Brass, Inducing Disbelief in Free Will Alters Brain Correlates of Preconscious Motor Preparation: The Brain Minds Whether We Believe in Free Will or Not. *Psychological Science* **22** (2011), 613-618.
[31] K. Vohs and J. Schooler. The Value of Believing in Free Will: Encouraging a Belief in Determinism Increases Cheating. *Psychological Science* **19** (2008): 49-54. . (doi:10.1111/j.1467-9280.2008.02045.x)
[32] R. Trivers, Deceit and self-deception: The relationship between communication and consciousness. In Robinson, M and Tiger, L. eds. Man and Beast Revisited, Smithsonian Press, Washington, DC, 1991
[33] M. Hauser, F. Cushman, L. Young, R. Kang-Xing Jin and J. Mikhail, A Dissociation Between Moral Judgments and Justifications. *Mind&Language* **22:1** (2007), 1-27
[34] M. Minsky, *The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind*, Simon & Schuster, New York, NY, 2006
[35] A. Bandura, Reconstrual of "free will" from the agentic perspective of social cognitive theory. In J. Baer, J. C. Kaufman, & R. F. Baumeister (Eds.), *Are we free?: Psychology and free will*, 86-127, Oxford University Press , Oxford, UK, 2008.