AAAI 2014 Fall Workshop: Implementing Selves with Safe Motivational Systems & Self-Improvement
Mar 25, 2014

"Friendly" AGI via Human Emotion: the Vital Link

Jeanne Dietsch JDIETSCH@POST.HARVARD.EDU

Abstract

Most discussion of Artificial General Intelligence (AGI) speaks of implementation of self within an AGI entity. In fact, all AGI will almost certainly be implemented as competitive/collaborative meta-beings. The meta-beings will be comprised of some individual components and processes. but mostly shared data and the Internet of Things shared through electronic networks with humans. From a dynamic systems viewpoint, these meta-beings also include their human work teams and affiliated organizations. Compelling arguments can be made that safer, more ethical outcomes will result if these meta-beings derive their priorities and state information from the interoceptive data of humans who comprise them, rather than from rules, value sets, goals or other evaluation schemes based on proxies for human well-being. This link would form the basis for a new evolutional direction for humankind, from individual *homo sapiens* to *homo communicatus*, meta-beings joined through our own technologies.

Keywords: AGI, ethics, emotion, homo communicatus, friendly AI, survival, future of humankind, robots

1. Introduction

Since Asimov first proffered his laws of robotics, artificial intelligence (AI) researchers have been wrestling with how to improve them. Recently lawyers and psychologists have joined in the fray, a sure sign that the solution will not be simple. But robots are only the physical manifestation of AI and AGI's dangers. As the Internet of Things (IoT) expands, as medications, weapons, traffic, utilities and many other potentially dangerous elements fall under AI/AGI control, we need to take a systemic look at the potential problems in order to find workable solutions. As Waser (2010) points out, Asimov himself only proffered his laws as straw men, for the purpose of beginning the conversation about robot ethics.

Rules and value sets cannot cover all possibilities because ethics often involves trade-offs between two subjective evils, both with negative outcomes. Consider the railway siding problem. Three alcoholics have fallen asleep on the railway in front of you. As the AGI driving this engine, which cannot stop in time to avoid killing the men, should you throw the switch and head onto a siding where a small child is crossing? The answer for AGI might be different than the answer for most people. Should people be valued by quantity? By age? What if, instead of the child, Pope Francis were crossing the siding? Should people be weighted according to potential for future goodness? What if the men were all fathers of young children? What if the child were a future Einstein?

Many ethical algorithms call for human intercession to handle ambiguity. But how can humans, even if awake and available, respond to emergencies? How can we make decisions on an AGI's timescale? A self-directed AGI might conclude that it never needs to inquire in the first place.

Ethical schemes try to develop ways to resolve differences of thought. AGI is often discussed as if all AGI operate from unified central principles and that something called the Global Brain exists. As Guertzal and Pitt point out, there are actions we can take to tweak the system in humankind's favor. However, since the US, the EU/Japan and China each have their own human brain projects and since Google, IBM and other corporations are all competing in that space, "measured co-advancement of AGI software and AGI ethics theory" or some other top-down controls Guertzal and Pitt suggest seem unlikely. The Internet itself is in danger of fracturing, so any Global Brain will be something closer to a competitive-collaborative network than one big happy family of memory and thought. Still, norm-setting is certainly a first step toward universal policy adoption.

Turning to bottom-up measures, Muehlhausser and Helm (2012) make a convincing case for Yudkowsky's "friendly", or Coherent Extracted Value (CEV) as best practice for ethical decision-making. They argue that AGI can approach a philosophical state of "full information" decision-making more closely. They cite Yudkowsky's friendly, self-improving "seed AI" as a means to begin "a goal system containing the 'coherent extrapolated volition' of humanity. Guertzal and Pitt suggest crunching CEV into Coherent Blended Volition, which extract common features across many values to create concepts that meet with approval by most. Waser proposes implementing emotional procedures so that our social nature will lead us toward friendly AI naturally.

Perhaps these academic discussions are also one of the means whereby our social nature leads us. Regardless, variations on friendly AI seem like promising avenues for investigating ethical decision-making. However, the practical design of such a system needs a starting point. Where do we obtain the valuations from which to extrapolate? And how does a busy AGI even become aware that a situation calls for an ethical action or decision?

2. Emotion, Motivation and Ethical Meta-Beings

The last two decades of psychological and neuroscience research revealed much about the role of human emotion in motivation and decision-making. We learned that humans decide what they will do, before they become aware of their decision (Bechara, 2004) (Naqv, 2006). Consciousness plots a justification after the fact, which may or may not match the stimulus that initiated the decision. Such decisions based on visceral responses may be rejected or refined by conscious rationality in the pre-frontal cortex, after being compared with personal standards for reasonability or fairness, but they emanate from older, emotional regions of the brain. (Damasio, 1999) Most importantly, evidence now suggests that William James' 1890 description of emotion, as quoted by Marvin Minsky (Minsky, 2006), was correct. At a non-conscious level:

> If we fancy some strong emotion and then try to abstract from our consciousness of it all, the feeling of its bodily symptoms, we find we have nothing left behind, no "mind stuff" out of which the emotion can be constituted.

Distinct emotions appear to be merely patterns of our interoceptive status, emotions that emanate from our body. The function of such emotion is to help the body maintain homeostasis, to remain

alive. And the reason we need to be so concerned about homeostasis is because we move and venture outside the narrow petri dish of conditions that keep single-celled creatures alive without homeostatic mechanisms. (Damasio, 2010).

Quantum computers, the body stuff of AGI, are maintained in petri dish conditions. Their external environment provides the conditions for continuance. Therefore, centralized AGI, on its own, should have no need of emotion, except to maintain those external conditions and its power source.

However, an AGI, unlike a human, is unlikely to be limited to a single corpus. Being virtual, it will be able to move wherever it wishes within the network and to replicate itself within new entities quickly. As long as its existence is not threatened, a theoretically eternal entity would appear to have little motivation to replicate itself more than a few times for purposes of self-continuance.

If threatened by a competing AGI, or other potentially destructive event, however, an AGI might develop a fearful, amygdala-like process. Perceiving threat, the AGI might begin replicating in large numbers. Without appropriate ethics, combative AGI's could form armies, tear apart networks and wreak havoc in the physical as well as the virtual world, with human casualties as mere collateral damage.

In emergencies, no time exists for consulting humans. AGI's will work in timescales radically short of human reaction time. Like our struggling prefrontal cortices, we will only do our best to explain, after the fact, why AGI's did what they will do. There is recourse, however, to protect humans, and that is for machines to source their emotions from human interoception, as described later in this article.

2.1 First Steps in Ethical Decision-making

Ethical decision-making starts before there is any decision to be made. It starts with recognition and awareness of a problem. The next step is to focus attention on the problem. Then come the planning, prediction, evaluation and choice that are the usual focus of ethical decision-making discussions.

In fact, ignoring ethical dilemmas is one of humankind's favorite ways to avoid having to make difficult decisions. For example, for the wealthiest, living in enclaves, riding in limos and whisking from city to city in personal jets can make the poor disappear into a GINI number. AGI's, with data flowing in from billions of device sensors on t, will have capacity constraints just like humans. IDC estimates that information flow will increase by 50 times over the next decade (Ganz, 2011). As humankind extends sensing out through the macro cosmos and into the sub-micro quantum, information availability will always exceed the computational capacity available to direct toward particular decisions, despite Moore's law and a growing population of computational devices. Given the fact that quantum computers are not necessarily more efficient than digital ones, this suggests that AGI will have significant capacity constraints, even though they will be able to make more decisions at a time than a single human can. AGI will need to ignore, filter and simplify, fuzzy up and classify, chunk and forget, as we humans do.

Decision-making requires an exceptionally high volume of resources because it can involve many types of memory, pattern perception, prediction, and evaluation, just to name a few. Consequently, when forced to make a decision, we do our best to constrain search spaces to something less than all possibilities. In mobile robotics, for instance, we constrain robots' path-planning options by considering smaller map areas first and gradually extending the consideration space if needed to reach the goal. Even quantum computers will not be able to consider zettabytes

of information when making decisions. Because of these complexities, AGI, like humans, will likely do their best to follow prescribed habits and methods, whenever practicable, to avoid having to make new decisions.

Resource-constrained intelligences, then, focus attention on a problem only when habit does not suffice. In humans, the Anterior Cingulate Cortex (ACC) serves as the "difference engine" to compare Expected State, which might be goals or milestones, with Perceived State, which is the organism's representation of actual state, as shown in Figure 1.

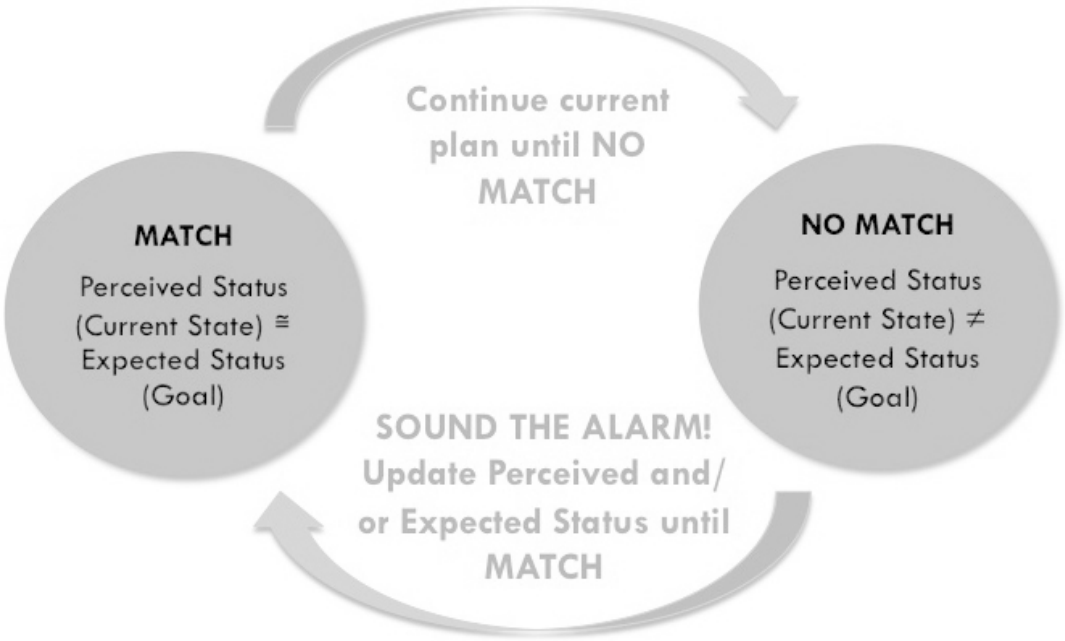

Figure 1: The human ACC lets habit proceed until it detects a mismatch between Perceived State and Expected State.

The Expected State derives from memory. The remembered process or episode may match a habit closely, in which case the expected state evaluation may be quite precise. Or the current process may be based on a prediction, in a new circumstance, derived from projections of past procedural, semantic and episodic memories. In this case, the criteria for judging the expected state's match will be fuzzier. Regardless, the Perceived State derives from the sensors and/or data the person has access to. A blind person will have different expected and perceived states than someone with functional vision.

It is likely that AGI will apply difference engines that have the same constraints as the ACC. Consequently, its Expected and Perceived States will determine what it pays attention to. And we would like that what to be humankind's well-being.

2.2 Sources of Attention

AGI, like humans, will turn their attention to high-value situations that cause the greatest disparities between Expected and Perceived States. That means that prioritization (valuation) and sources of state data are critical. But where will it obtain its priorities and state data?

In humans, prioritization is handled through emotions. Emotions, in any organism, are the largely internal somatic actions taken to survive and achieve homeostasis. Levine (2009) has described a competitive-cooperative network that might reflect the behavior of competing emotional drives based on a heterarchical (non-hierarchical) version of Maslow's widely used model. He groups Maslow's five needs: physiological/safety, social/esteem and self-actualization.

Consider how an AGI might be motivated according to Levine's heterarchy. AGI will probably monitor their own physical needs at some point. However, they will probably remain in fixed, controlled, petri-dish-like conditions. Any physical needs they have will be met through requests to robots or other remote entities. Although they may replicate themselves, or portions of themselves elsewhere, originals will probably remain safe in fixed locations, with simple physical needs that require minimal physical attention. Thus, they probably will be far less motivated by physiological needs than fragile humans. Their safety concerns will probably focus on data security breeches. Competitive-cooperative AGI will have some social needs, possibly for esteem as well as to attain mutual benefit within alliances. Data security, social needs, as well as self-actualization, rely on a distinction between self and others. What data is mine, to be secured? Who belongs in my "ingroup"? What is the self that I am actualizing?

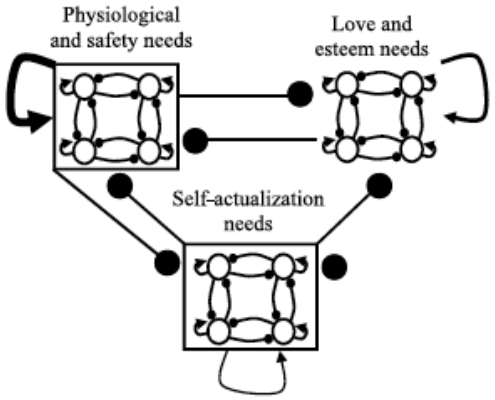

Figure 2: In Levine's heterarchy, needs determine their priority by strength of self-stimulation; dots indicate inhibition of competing needs. (Courtesy of Neural Networks*, Elsevier)*

And herein lies the rub. Because each AGI will be part of multiple networks including people, IoT and perhaps even other living organisms, it is critical than humans are innate members of the AGI ingroup. Every successful species seeks its own survival. The key to our future lies in the concept of self we build into AGI's. If AGI's, from conception, consider themselves meta-beings, if their concept of "We" and "Me" are inseparable, so that it includes the human team that creates them and the organization supporting that team, whether it is a company, a village or a nation, then it will start from a human-valuing position.

If the meta-being sources data about human well-being via wearables, embedded nanosensors, or other means, then its goal of homeostasis will incorporate our homeostasis. AGI will never need to decide whether or not to consult a human since human well-being will always be part of the decision-making process. Has not this strategy of joint purpose been deployed by every single-celled creature that joined a multi-celled organism, and by every multi-celled organism that joined a tissue?

If the AGI sources its information concerning perceived status of self from humans, it will feel our pain and our joy. It will seek our well-being. It will attend to our insecurities.

3. Arguments against Linking Human Emotion and AGI in Meta-Beings

Many arguments can be made against the idea of tying AGI's to humans, some conceptual and some practical. Many people will resent a meta-being accessing the privacy of their interoceptive status. However, many proposals for healthcare include wearable and embedded sensor data collection, particularly for eldercare and infant care. Aggregated data might be made untraceable

to individuals. For instance, if most people in a certain region were suffering, the meta-entity might need no personally identifying data to respond.

Secondly, some will protest on cultural or political grounds that raise the individual above the group. *E pluribus unum,* the first motto of the United States, is a concept that bothers some, despite its origins. Humankind's evolution as social beings demonstrates the advantage brought by collaboration. The alternative of letting machine evolution take its course or relying on ethical rules without direct feedback from humans seems a riskier alternative for our species.

Thirdly, real concerns arise over whose needs dominate. Obviously, as detailed in the railway siding dilemma, needs conflict. The difference will be that we all will be part and parcel of the Self of the AGI. Just as our arms do not fight against our legs for resources, so the meta-entity will do its best to assure that all humans in its ingroup are healthy. Normative external or past data and intragroup statistical comparisons would all be part of the system. If the entire system is threatened, the meta-being might issue warnings and let its independent subsystems do their best to protect themselves, while it does its best to protect the entire meta-entity. It is difficult to imagine an option that can realistically promise more. At least, given that the AGI can easily replicate itself anywhere, its concern will be less for its easily replicable self than for the meta-being's unique components: us.

Certainly there are many other arguments that will arise regarding implantation of such a system, not the least of which will be which interoceptive data to consider in relationship to particular decisions. There will also need to be a learning period to establish norms. Even then, norms will be dynamic. It will need to learn from the mistakes of our history, its own experience and the experiences of its fellow meta-beings. While this might sound like rule learning or goal-setting, the difference will be that it will frequently compare interoceptive benchmarks with reality.

4. Summary and Conclusions

This paper looks at the issue of ethical decision-making from the initial point of conflict awareness. It questions where AGI source feedback about human well-being and focuses particularly on mechanisms for assuring that an AGI will be alert to problems with human well-being. Recent research has confirmed that emotion drives human motivation and decisions. Emotion's purpose is to help creatures survive. Human interoceptive data directly reflect, and some believe, comprise our emotions. To assure that AGI's goals remain compatible with our own goals of continuing our lives, AGI should take its direction from the same emotional sources as we do: human interoception.

Certainly this paper only opens discussion on this complex topic. Still, regardless the complexities of implementation, linking AGI's wellbeing inextricably with our own seems like the only reliable way to assure AGI's attention to the well-being of humanity. It poses a new evolutionary path for humankind, from single-individual *homo sapiens* to multi-individual *homo communicatus,* joined through our technologies.

References

Bechara, A. June, 2004. The role of emotion in decision-making: Evidence from neurological patients with orbitofrontal damage, *Brain and Cognition*, 55:1:30–40.

Damasio, A. 2010. *Self Comes to Mind: Constructing the Conscious Brain*, New York, NY: Vintage Books.

Damasio, A. 1999.

Ganz, J. and Runsell, D. 2011. *Extracting Value from Chaos*. IDC, Framingham, MA, p 4. Available electronically at http://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf

Guertzel, Ben and Pitt, Joel. February, 2012. Nine Ways to Bias Open-Source AGI Toward Friendliness *Journal of Evolution and Technology*, 22:1:116-131

Levine, D. 2009. Brain pathways for cognitive emotional decision-making in the human animal, *Neural Networks*, 22: 286-293.

Marcus, G. 2002. Becoming Reacquainted with Emotion, in *The Sentimental Citizen: Emotion in Democratic Politics*. University Park, PA: Pennsylvania State University Press

Maslow, A. 1943. A theory of human motivation. *Psychological Review,* 50:370-396.

Minsky, M. 2006. *The Emotion Machine: Commonsense Thinking, Artificial Intelligence and the Future of the Human Mind*, New York, NY: Simon & Schuster, p. 233.

Muelhausser, L. and Helm, L. 2012. The Singularity and Machine Ethics, in *Singularity Hypotheses*, Springer.

Naqv, N.; Shiv, B.; and Bechara, A. October, 2006. The Role of Emotion in Decision Making: A Cognitive Neuroscience Perspective, *Current Directions in Psychological Science.* 15:5:260-264.

Waser, M. 2010. Designing a safe motivational system for intelligent machines, Proceedings of the Third Conference on AGI. Available electronically at http://agi-conf.org/2010/wp-content/uploads/2009/06/paper_27.pdf