Journal of EVOLUTION & Technology

A peer-reviewed electronic journal published by the Institute for Ethics and Emerging Technologies

ISSN 1541-0099

22(1) – February 2012

Nine Ways to Bias Open-Source AGI Toward Friendliness

Ben Goertzel and Joel Pitt
Novamente LLC
ben@goertzel.org

Journal of Evolution and Technology - Vol. 22 Issue 1 – February 2012 - pgs 116-131

Abstract

While it seems unlikely that any method of *guaranteeing* human-friendliness ("Friendliness") on the part of advanced Artificial General Intelligence (AGI) systems will be possible, this doesn't mean the only alternatives are throttling AGI development to safeguard humanity, or plunging recklessly into the complete unknown. Without denying the presence of a certain irreducible uncertainty in such matters, it is still sensible to explore ways of *biasing the odds* in a favorable way, such that newly created AI systems are significantly more likely than not to be Friendly. Several potential methods of effecting such biasing are explored here, with a particular but non-exclusive focus on those that are relevant to open-source AGI projects, and with illustrative examples drawn from the OpenCog open-source AGI project. Issues regarding the relative safety of open versus closed approaches to AGI are discussed and then nine techniques for biasing AGIs in favor of Friendliness are presented:

1. Engineer the capability to acquire integrated ethical knowledge.
2. Provide rich ethical interaction and instruction, respecting developmental stages.
3. Develop stable, hierarchical goal systems.
4. Ensure that the early stages of recursive self-improvement occur relatively slowly and with rich human involvement.
5. Tightly link AGI with the Global Brain.
6. Foster deep, consensus-building interactions between divergent viewpoints.
7. Create a mutually supportive community of AGIs.
8. Encourage measured co-advancement of AGI software and AGI ethics theory
9. Develop advanced AGI sooner not later.

In conclusion, and related to the final point, we advise the serious co-evolution of functional AGI systems and AGI-related ethical theory as soon as possible, before we have so much technical infrastructure that parties relatively unconcerned with ethics are able to rush ahead with brute force approaches to AGI development.

1. Introduction

Artificial General Intelligence (AGI), like any technology, carries both risks and rewards. One science fiction film after another has highlighted the potential dangers of AGI, lodging the issue deep in our cultural awareness. Hypothetically, an AGI with superhuman intelligence and capability could dispense with humanity altogether and thus pose an "existential risk" (Bostrom 2002). In the worst case, an evil but brilliant AGI, programmed by some cyber Marquis de Sade, could consign humanity to unimaginable tortures (perhaps realizing a modern version of the medieval Christian imagery of hell). On the other hand, the potential benefits of powerful AGI also go literally beyond human imagination. An AGI with massively superhuman intelligence and a positive disposition toward humanity could provide us with truly dramatic benefits, through the application of superior intellect to scientific and engineering challenges that befuddle us today. Such benefits could include a virtual end to material scarcity via advancement of molecular manufacturing, and also force us to revise our assumptions about the inevitability of disease and aging (Drexler1986). Advanced AGI could also help individual humans grow in a variety of directions, including directions leading beyond our biological legacy, leading to massive diversity in human experience, and hopefully a simultaneous enhanced capacity for openmindedness and empathy.

Eliezer Yudkowsky introduced the term "Friendly AI" to refer to advanced AGI systems that act with human benefit in mind (Yudkowsky 2001). Exactly what this means has not been specified precisely, though informal interpretations abound. Goertzel (2006a) has sought to clarify the notion in terms of three core values of "Joy, Growth and Freedom." In this view, a Friendly AI would be one that advocates individual and collective human joy and growth, while respecting the autonomy of human choice.

Some (for example, De Garis 2005) have argued that Friendly AI is essentially an impossibility, in the sense that the odds of a dramatically superhumanly intelligent mind worrying about human benefit are vanishingly small, drawing parallels with humanity's own exploitation of less intelligent systems. Indeed, in our daily life, questions such as the nature of consciousness in animals, plants, and larger ecological systems are generally considered merely philosophical, and only rarely lead to individuals making changes in outlook, lifestyle or diet. If Friendly AI is impossible for this reason, then the best options for the human race would presumably be to avoid advanced AGI development altogether, or else to fuse with AGI before the disparity between its intelligence and humanity's becomes too large, so that beings-originated-as-humans can enjoy the benefits of greater intelligence and capability. Some may consider sacrificing their humanity an undesirable cost. The concept of humanity, however, is not a stationary one, and can be viewed as sacrificed from only our contemporary perspective of what humanity is. With our cell phones, massively connected world, and the inability to hunt, it's unlikely that we'd seem the same species to the humanity of the past. Just like an individual's self, the self of humanity will inevitably change, and as we do not usually mourn losing our identity of a decade ago to our current self, our current concern for what we may lose may seem unfounded in retrospect.

Others, such as Waser (2008), have argued that Friendly AI is essentially inevitable, linking greater intelligence with greater cooperation. Waser adduces evidence from evolutionary and human history in favor of this point, along with more abstract arguments such as the economic viability of cooperation over not cooperating.

Omohundro (2008) has argued that any advanced AI system will very likely demonstrate certain "basic AI drives," such as desiring to be rational, to self-protect, to acquire resources, and to preserve and protect its utility function and avoid counterfeit utility; these drives, he suggests, must be taken carefully into account in formulating approaches to Friendly AI.

Yudkowsky (2006) discusses the possibility of creating AGI architectures that are in some sense "provably Friendly" – either mathematically, or else by very tight lines of rational verbal argument. However, several possibly insurmountable challenges face such an approach. First, proving mathematical results of this nature would likely require dramatic advances in multiple branches of mathematics. Second,

such a proof would require a formalization of the goal of "Friendliness," which is a subtler matter than it might seem (Legg 2006; Legg 2006a), as formalization of human morality has vexed moral philosophers for quite some time. Finally, it is unclear the extent to which such a proof could be created in a generic, environment-independent way – but if the proof depends on properties of the physical environment, then it would require a formalization of the environment itself, which runs up against various problems related to the complexity of the physical world, not to mention the current lack of a complete, consistent theory of physics.

The problem of formally or at least very carefully defining the goal of Friendliness has been considered from a variety of perspectives. Among a list of fourteen objections to the Friendly AI concept, with suggested answers to each, Sotala (2011) includes the issue of friendliness being a vague concept. A primary contender for this role is the concept of Coherent Extrapolated Volition (CEV) suggested by Yudkowsky (2004), which roughly equates to the extrapolation of the common values shared by all people when at their best. Many subtleties arise in specifying this concept – e.g. if Bob Jones is often possessed by a strong desire to kill all Martians, but he deeply aspires to be a nonviolent person, then the CEV approach would not rate "killing Martians" as part of Bob's contribution to the CEV of humanity. Resolving inconsistencies in aspirations and desires, and the different temporal scales involved for each, is another non-trivial problem.

One of the authors, Goertzel (2010), has proposed a related notion of Coherent Aggregated Volition (CAV), which eschews some subtleties of extrapolation, and instead seeks a reasonably *compact*, *coherent*, and *consistent* set of values that is close to the collective value-set of humanity. In the CAV approach, "killing Martians" would be removed from humanity's collective value-set because it's assumedly uncommon and not part of the most compact/coherent/consistent overall model of human values, rather than because of Bob Jones's aspiration to nonviolence.

More recently we have considered that the core concept underlying CAV might be better thought of as CBV or Coherent Blended Volition. CAV seems to be easily misinterpreted as meaning the average of different views, which was not the original intention. The CBV terminology clarifies that the CBV of a diverse group of people should not be thought of as an average of their perspectives, but as something more analogous to a "conceptual blend" (Fauconnier and Turner 2002) – incorporating the most essential elements of their divergent views into a whole that is overall compact, elegant and harmonious. The subtlety here (to which we shall return below) is that for a CBV blend to be broadly acceptable, the different parties whose views are being blended must agree to some extent that enough of the essential elements of their own views have been included.

Multiple attempts at axiomatization of human values have also been attempted. In one case, this is done with a view toward providing near-term guidance to military robots (from Arkin (2009)'s excellent though chillingly-titled book *Governing Lethal Behavior in Autonomous Robots*). However, there are reasonably strong arguments that human values (and similarly a human's language and perceptual classification) are too complex and multifaceted to be captured in any compact set of formal logical rules. Wallach and Allen (2010) have made this point eloquently, and argued for the necessity of fusing top-down (e.g. formal logic based) and bottom-up (e.g. self-organizing learning based) approaches to machine ethics.

1.1 Modes of AGI development

Other sociological considerations also arise. For example, it is sometimes argued that the risk from highly-advanced AGI going morally awry on its own may be less than that of moderately-advanced AGI being used by a human being to advocate immoral ends. This possibility gives rise to questions about the ethical value of various practical paths of AGI development, for instance:

- Should AGI be developed in a top-secret installation by a select group of individuals? Individuals selected for a combination of technical and scientific brilliance, moral uprightness, or any other

qualities deemed relevant (a "closed approach")? Or should it be developed in the open, in the manner of open-source software projects like Linux (an "open approach")? The open approach allows the collective intelligence of the world to participate more fully – but also potentially allows unscrupulous elements of the human race to take some of the publicly-developed AGI concepts and tools, then privately develop them into AGIs with selfish or evil purposes in mind. Is there some meaningful intermediary between these extremes?

- Should governments regulate AGI, with Friendliness in mind (as advocated carefully by e.g Hibbard (2002))? Or will this just cause AGI development to move to the handful of countries with more liberal policies? Or will it cause development to move underground, where nobody can see the dangers developing?

Clearly, there are many subtle and interwoven issues at play here, and it may take an AGI beyond human intelligence to unravel and understand them all thoroughly. Our goal here is more modest: to explore the question of *how to militate in favor of positive, Friendly outcomes*. Some of our suggestions are fairly generic, but others are reliant on the assumption of an open rather than closed approach. The open approach is currently followed in our own AGI project, hence its properties are those we're most keen to explore.

While we would love to be proven wrong on this, our current perspective is that provably, or otherwise guarantee-ably, Friendly AI is not achievable. On the face of it, achieving strong certainty about the future behaviors of beings massively more generally intelligent and capable than ourselves seems implausible. Again, we are aiming at a more modest goal – to explore ways of biasing the odds, and creating AI systems that are significantly more likely than not to be Friendly.

While the considerations presented here are conceptually fairly generic, we will frequently elaborate them using the example of the OpenCog (Goertzel et al. 2010; Hart and Goertzel 2008) AGI framework on which we are currently working, and the specific OpenCog applications now under development, including game AI, robotics, and natural language conversation.

2. Is open or closed AGI development safer?

We will not seek here to argue rigorously that the open approach to AGI is preferable to the closed approach. Rather, our goal is to explore ways to make AGI more probably Friendly, with a non-exclusive focus on open approaches. We do believe intuitively that the open approach is probably preferable, but our reasons are qualitative and we recognize there are also qualitative arguments in the opposite direction. Before proceeding further, we will briefly sketch some of the reasons for our intuition on this.

First, we have a strong skepticism about self-appointed elite groups that claim that they know what's best for everyone (even if they are genuine saints), and a healthy respect for the power of collective intelligence and the Global Brain (Heylighen 2007), which the open approach is ideal for tapping. On the other hand, we also understand the risk of terrorist groups or other malevolent agents forking an open source AGI project and creating something terribly dangerous and destructive. Balancing these factors against each other rigorously is impossible, due to the number of assumptions currently involved.

For instance, nobody really understands the social dynamics by which open technological knowledge plays out in our current world, let alone hypothetical future scenarios. Right now there exists open knowledge about many very dangerous technologies, and there exist many terrorist groups, yet these groups fortunately make scant use of the technologies. The reasons appear to be essentially sociological – the people involved in terrorist groups tend not to be the ones who have mastered the skills of turning public knowledge of cutting-edge technologies into real engineered systems. While it's easy to observe this sociological phenomenon, we certainly have no way to estimate its quantitative extent from first principles. We don't really have a strong understanding of how safe we are right now, given the

technological knowledge available via the Internet, textbooks, and so forth. Relatively straightforward threats such as nuclear proliferation remain confusing, even to the experts.

The open approach allows for various benefits of open source software development to be applied, such as Linus's law (Raymond 2000): "Given enough eyeballs, all bugs are shallow."

Software development practice has taught us that in the closed approach it's very hard to get the same level of critique as one obtains on a public, open codebase. At a conceptual level of development, a closed approach also avoids making it possible for external theorists to find specific flaws in a design. Discussing the theoretical basis for Friendliness design is all very well, but implementing and designing a system that conforms to that design is another.

Keeping powerful AGI and its development locked up by an elite group doesn't really provide reliable protection against malevolent human agents either. History is rife with such situations going awry, such as the leadership of the group being subverted, brute force being inflicted by some outside party, or a member of the elite group defecting to some outside group in the interest of personal power, reward, or internal group disagreements. There are many things that can go wrong in such situations, and the confidence of any particular group that it is immune to such issues, cannot be taken very seriously.

Clearly, neither the open nor closed approach qualifies as a panacea.

3. The (unlikely) prospect of government controls on AGI development

Given the obvious long-term risks associated with AGI development, is it feasible that governments might enact legislation intended to stop AI from being developed? Surely government regulatory bodies would slow down the progress of AGI development in order to enable measured development of accompanying ethical tools, practices, and understandings? This however seems unlikely, for the following reasons.

Let us consider two cases separately. First, there is the case of banning AGI research and development *after* an "AGI Sputnik" moment has occurred. We define an AGI Sputnik moment as a technological achievement that makes the short- to medium-term possibility of highly functional and useful human-level AGI broadly evident to the public and policy makers, bringing it out of the realm of science fiction to reality. Second, we might choose to ban it *before* such a moment has happened.

After an AGI Sputnik moment, even if some nations chose to ban AI technology due to the perceived risks, others would probably proceed eagerly with AGI development because of the wide-ranging perceived benefits. International agreements are difficult to reach and enforce, even for extremely obvious threats like nuclear weapons and pollution, so it's hard to envision that such agreements would come rapidly in the case of AGI. In a scenario where some nations ban AGI while others do not, it seems the slow speed of international negotiations would contrast with the rapid speed of development of a technology in the midst of revolutionary breakthrough. While worried politicians sought to negotiate agreements, AGI development would continue, and nations would gain increasing competitive advantage from their differential participation in it.

The only way it seems feasible for such an international ban to come into play, would be if the AGI Sputnik moment turned out to be largely illusory because the path from the moment to full human-level AGI turned out to be susceptible to severe technical bottlenecks. If AGI development somehow slowed after the AGI Sputnik moment, then there might be time for the international community to set up a system of international treaties similar to what we now have to control nuclear weapons research. However, we note that the nuclear weapons research ban is not entirely successful – and that nuclear weapons development and testing tend to have large physical impacts that are remotely observable by foreign nations. On the other hand, if a nation decides not to cooperate with an international AGI ban, this would be much more difficult for competing nations to discover.

An unsuccessful attempt to ban AGI research and development could end up being far riskier than no ban. An international R&D ban that was systematically violated in the manner of current international nuclear weapons bans would shift AGI development from cooperating developed nations to "rogue nations," thus slowing down AGI development somewhat, but also perhaps decreasing the odds of the first AGI being developed in a manner that is concerned with ethics and Friendly AI.

Thus, subsequent to an AGI Sputnik moment, the overall value of AGI will be too obvious for AGI to be effectively banned, and monitoring AGI development would be next to impossible.

The second option is an AGI R&D ban earlier than the AGI Sputnik moment – before it's too late. This also seems infeasible, for the following reasons:

- Early stage AGI technology will supply humanity with dramatic economic and quality of life improvements, as narrow AI does now. Distinguishing narrow AI from AGI from a government policy perspective would also be prohibitively difficult.

- If one nation chose to enforce such a slowdown as a matter of policy, the odds seem very high that other nations would explicitly seek to accelerate their own progress on AI/AGI, so as to reap the ensuing differential economic benefits.

To make the point more directly, the prospect of any modern government seeking to put a damper on current real-world narrow-AI technology seems remote and absurd. It's hard to imagine the US government forcing a roll-back from modern search engines like Google and Bing to more simplistic search engines like 1997 AltaVista, on the basis that the former embody natural language processing technology that represents a step along the path to powerful AGI.

Wall Street firms (that currently have powerful economic influence on the US government) will not wish to give up their AI-based trading systems, at least not while their counterparts in other countries are using such systems to compete with them on the international currency futures market. Assuming the government did somehow ban AI-based trading systems, how would this be enforced? Would a programmer at a hedge fund be stopped from inserting some more-effective machine learning code in place of the government-sanctioned linear regression code? The US military will not give up their AI-based planning and scheduling systems, as otherwise they would be unable to utilize their military resources effectively. The idea of the government placing an IQ limit on the AI characters in video games, out of fear that these characters might one day become too smart, also seems absurd. Even if the government did so, hackers worldwide would still be drawn to release "mods" for their own smart AIs inserted illicitly into games; and one might see a subculture of pirate games with illegally smart AI.

"Okay, but all these examples are narrow AI, not AGI!" you may argue. "Banning AI that occurs embedded inside practical products is one thing; banning autonomous AGI systems with their own motivations and self-autonomy and the ability to take over the world and kill all humans is quite another!" Note though that the professional AI community does not yet draw a clear border between narrow AI and AGI. While we do believe there is a clear qualitative conceptual distinction, we would find it hard to embody this distinction in a rigorous test for distinguishing narrow AI systems from "proto-AGI systems" representing dramatic partial progress toward human-level AGI. At precisely what level of intelligence would you propose to ban a conversational natural language search interface, an automated call center chatbot, or a house-cleaning robot? How would you distinguish rigorously, across all areas of application, a competent non-threatening narrow-AI system from something with sufficient general intelligence to count as part of the path to dangerous AGI?

A recent workshop of a dozen AGI experts, oriented largely toward originating such tests, failed to come to any definitive conclusions (Adams et al. 2010), recommending instead that a looser mode of evaluation

be adopted, involving qualitative synthesis of multiple rigorous evaluations obtained in multiple distinct scenarios. A previous workshop with a similar theme, funded by the US Naval Research Office, came to even less distinct conclusions (Laird et al. 2009). The OpenCog system is explicitly focused on AGI rather than narrow AI, but its various learning modules are also applicable as narrow AI systems, and some of them have largely been developed in this context. In short, there's no rule for distinguishing narrow AI work from proto-AGI work that is sufficiently clear to be enshrined in government policy, and the banning of narrow AI work seems infeasible as the latter is economically and humanistically valuable, tightly interwoven with nearly all aspects of the economy, and nearly always non-threatening in nature. Even in the military context, the biggest use of AI is in relatively harmless-sounding contexts such as back-end logistics systems, not in frightening applications like killer robots.

Surveying history, one struggles to find good examples of advanced, developed economies slowing down development of any technology with a nebulous definition, obvious wide-ranging short to medium term economic benefits, and rich penetration into multiple industry sectors, due to reasons of speculative perceived long-term risks. Nuclear power research is an example where government policy has slowed things down, but here the perceived economic benefit is relatively modest, the technology is restricted to one sector, the definition of what's being banned is very clear, and the risks are immediate rather than speculative. More worryingly, nuclear weapons research and development continued unabated for years, despite the clear threat it posed.

In summary, we submit that, due to various aspects of the particular nature of AGI and its relation to other technologies and social institutions, it is very unlikely to be explicitly banned, either before or after an AGI Sputnik moment. If one believes the creation of AGI to be technically feasible, then the more pragmatically interesting topic becomes how to most effectively manage and guide its development.

4. Nine ways to bias AGI toward Friendliness

There is no way to guarantee that advanced AGI, once created and released into the world, will behave according to human ethical standards. There is irreducible risk here, and in a sense it is a risk that humanity has been moving towards, at accelerating speed, ever since the development of tools, language, and culture. However, there are things we can do to bias the odds in the favor of ethically positive AGI development. The degree of biasing that can be achieved seems impossible to estimate quantitatively, and any extrapolation from human history to a future populated by agents with significantly transhuman general intelligence has an obvious strong risk of being profoundly flawed. Nevertheless, it behooves us to do our best to bias the outcome in a positive direction, and the primary objective of this paper is to suggest some potential ways to do so.

4.1 Engineer the capability to acquire integrated ethical knowledge

First of all, if we wish our AGI systems to behave in accordance with human ethics, we should design them to be capable of the full range of human ethical understanding and response. As reviewed in Goertzel and Bugaj (2008) and Goertzel (2009b), human ethical judgment relies on the coordination and integration of multiple faculties. One way to think about this is to draw connections between the multiple types of human memory (as studied in cognitive psychology and cognitive neuroscience) and multiple types of ethical knowledge and understanding. To wit:

- **Episodic memory** corresponds to the process of ethically assessing a situation based on similar prior situations.

- **Sensorimotor memory** corresponds to "mirror neuron" (Rizzolatti and Craighero 2004) type ethics, where you feel another person's feelings via mirroring their physiological emotional responses and actions.

- **Declarative memory** corresponds to rational ethical judgment.

- **Procedural memory** corresponds to "ethical habit": learning by imitation and reinforcement to do what is right, even when the reasons aren't well articulated or understood.

- **Attentional memory** corresponds to the existence of appropriate patterns guiding one to pay adequate attention to ethical considerations at appropriate times.

- **Intentional memory** corresponds to ethical management of one's own goals and motivations (e.g. when do the ends justify the means?).

We argue that an ethically mature mind, human or AGI, should balance all these kinds of ethics, although none is completely independent of the others.

How these memory types relate to ethical behavior and understanding depends somewhat on the cognitive architecture in question. For instance, it is straightforward to identify each of these memory types in the OpenCog architecture, and articulate therein their intuitive relationship to ethical behavior and understanding:

- **Episodic memory**: Through placing OpenCog in ethical scenarios with a teacher agent that provides feedback on choices, and with OpenCog's goal system initially biased to seek approval from the teacher.

- **Sensorimotor memory**: Knowledge is usually contextually represented within the OpenCog AtomSpace (a weighted hypergraph-like knowledge base). A perceptual interface that takes on the role of mirror neurons may activate contexts representing another's emotional state, causing that context to move into the attentional focus of OpenCog. In this way, OpenCog becomes sensitive to the emotional state of other agents it has interacted with and modelled the world view of. Then, through induction or pattern mining these changes in emotional state can be mapped on to new agents that the AI is unfamiliar with.

- **Declarative memory**: Declarative ethical knowledge may be embedded as a seed within the OpenCog AtomSpace, or built from data mining episodic memory for patterns learned during ethical teaching. This knowledge can then be reasoned about, using probablistic logic to make ethical decisions in novel situations.

- **Procedural memory**: The development of new schema can be based on previous experience. Schema that have previously been evaluated in the same or similar ethical scenarios can be used to guide the construction of new program trees.

- **Attentional memory**: OpenCog has networks of attention that can implicitly store attentional memories. These memories form from observation of temporal patterns of knowledge access, and their relative importance to goal fulfillment. Once formed they degrade slowly and may provide resilience against potentially unethical replacements if initially taught ethical behavior (Goertzel et al. 2010).

- **Intentional memory** (memory regarding goals and subgoals): OpenCog expresses explicit goals declaratively using uncertain logic, but also expresses implicit goals using "maps" recording habitual patterns of activity, created and stored via attentional memory.

Also worth noting in this context is the theory of "Stages of Ethical Development in Artificial General Intelligence Systems" presented in Goertzel and Bugaj (2008). This theory integrates, among other aspects, Kohlberg's (1981) theory of logical ethical judgment (focused on justice and declarative

knowledge) and Gilligan's (1982) theory of empathic ethical judgment (focused on interpersonal relationships formed from episodic and sensorimotor memory). In this integrated theory, as shown in Tables 1, 2, and 3 (see Appendix), it is asserted that, to pass beyond childish ethics to the "mature" stage of ethical development, a deep and rich integration of the logical and empathic approaches to ethics is required. Here we suggest a slight modification to this idea: to pass to the mature stage of ethical development, a deep and rich integration of the ethical approaches associated with the six main types of memory systems is required. Of course, there are likely to be other valuable perspectives founded on different cognitive models, and this is an area wide open for further exploration both conceptually and empirically.

4.2 Provide rich ethical interaction and instruction, respecting developmental stages

Of course, a cognitive architecture with capability to exercise the full richness of human ethical behavior and understanding is not enough – there next arises the question of how to fill this cognitive architecture with appropriate "ethical content." Just as human ethics are considered a combination of nature and nurture, so we should expect for AGI systems. AGI systems are learning systems by definition, and human values are complex and best conveyed via a combination of methods in order that they become well grounded.

In Goertzel (2009a) the memory types listed in the previous section are associated with different common modes of human communication:

Memory Type	**Communication Type**	**Description**
sensorimotor	depictive	in which an agent creates some sort of (visual, auditory, etc.) construction to show another agent, with a goal of causing the other agent to experience phenomena similar to what they would experience upon experiencing some particular entity in the shared environment
episodic	dramatic	in which an agent creates an evocation of specific scenes or episodes in which to evoke particular real or imagined episodes in the other agent's mind
declarative	linguistic	communication using language whose semantics are largely (not necessarily wholly) interpretable based on the mutually experienced world
procedural	demonstrative	in which an agent carries out a set of actions in the world, and the other agent is able to imitate these actions, or instruct another agent as to how to imitate these actions
attentional	indicative	in which e.g. one agent points to some part of the world or delimits some interval of time, and another agent is able to interpret the meaning
intentional	intentional	in which an agent explicitly communicates to another agent what its goal is in a certain situation (in humans this relates closely to mirror neuron activity: Rizzolatti and Craighero 2004)

Our suggestion is that AGIs should be provided with ample ethical instruction using all of the above communication modalities. During this instruction, respect for modern thinking about progressive education will be important. Among this thinking is that it is important to tailor ethical instruction to the student's stage of cognitive and ethical development. Instructions on the abstract nature of justice are not likely to be helpful to an AGI that hasn't yet learned the practicalities of sharing with its peers – at that early stage, abstract ethical instructions would constitute ungrounded declarative knowledge, and the AGI

system would have a hard time grounding them and integrating them with its overall world view and express it in all the different forms of memory available to it. Whereas after an AGI has learned some of the everyday aspects of justice, including the balance of justice with empathy in everyday life, and once it has also gotten familiar with the application of abstract principles to other aspects of ordinary life, it will be well poised to appreciate abstract ethical principles and their utility in making difficult decisions – it will be able to understand the abstract nature of justice in a richer and more holistic way.

More concretely, to make just a few obvious points:

1. The teacher(s) should be observed to follow their own ethical principles, in a variety of contexts that are meaningful to the AGI. Without it, declarative memory may clash with episodic (or other memory types). However, at the same time, perceived inconsistencies in the behavior of the teacher may hint at subtleties in human ethics which the AGI was not previously aware of. In such a case, questioning the teacher on this discrepancy may refine the AGI's understanding.

2. The system of ethics must be relevant to the AGI's life context and embedded within its understanding of the world. Without this, episodic memories may not be sufficiently similar to new situations to engage an ethical action or response when it should.

3. Ethical principles must be grounded in both theory-of-mind thought experiments (emphasizing logical coherence) and real-life situations in which the ethical trainee is required to make a moral judgment and is rewarded or reproached by the teacher(s). The feedback should also include imparting explanatory augmentations to the teachings regarding the reason for a particular decision on the part of the teacher.

For example, in our current application of OpenCog to control intelligent game characters, we intend to have human players take the role of the teacher in a shared sandbox environment. The AGI can not only interact with the teacher through dialogue and action, but can also observe the teacher interacting with other humans and AGIs, including how they are rewarded or chastised. Initially, teaching should occur for each embodiment option: each game world in which an AGI has an virtual avatar, and each robotic body available to the AGI. Eventually, a sufficient corpus of varied episodic knowledge will allow the AGI to extract commonalities between embodied instances; which, in turn, will encourage commensurability.

4.3 Create stable, hierarchy-dominated goal systems

One aspect of cognitive architecture is especially closely associated with ethical issues: goals and motivations. This is an area where, we suggest, the best path to creating highly ethical AGI systems may be to deviate from human cognitive architecture somewhat.

Some may perceive this as a risky assertion – since, after all, the human cognitive architecture is moderately well understood, whereas any new direction will bring with it additional uncertainties. However, the ethical weaknesses of the human cognitive architecture are also very well understood, and we see no reason to believe that seeking to implement a closely human-like goal system and ethical system in an AGI system that differs from humans in significant ways (e.g. a robot body rather than a human body, no mother and father, no rhythmic breathing flow giving it innate empathy with the rhythms of nature, etc.) would yield predictable or positive results. Indeed, if we could really create a digital human, correct down to a fairly detailed level, with a closely human-like body, then we would have a system whose ethical behavior and thinking we would be able to understand very well by analogy to ordinary humans. We might find this digital human to possess profound psychological and ethical difficulties due to its lack of an ordinary biological heritage and family, and then we could try to deal with these issues using tools of human psychology and psychiatry. Or we might even choose to implant such a digital human with false memories of a human heritage and family, and experiment with the ethically questionable consequences.

Apart from scenarios like these, however, if we're talking about taking a human-like AGI mind and embodying it in a video-game world, a plastic/metal humanoid robot body, or only a text chat interface, we are already talking about a system operating in a regime very different from any historical human being. For instance, empathy is a very important component of human ethics, and the roots of human empathy lie in our tactile relationship with our parents in our infancy, our ability to synchronize our breathing with the other humans around us, and a host of other aspects of our particular human embodiment. In taking a closely human-like cognitive architecture and lifting it out of the context of all this bodily intelligence, one is already doing something quite "artificial." So, barring a true mind-and-body digital-human approach (which seems infeasible in the short or medium term future), the choice is not human vs. non-human, but rather between different ways of constructing non-human-like AGIs by incorporating aspects of human architecture with engineered structures and dynamics. Given this reality, our considered opinion is that the approach most likely to yield an ethically positive outcome is to deviate significantly from the "intentional" component of human cognitive architecture, and create AGI systems embodying a different approach to goals and motivations. Specifically, we believe it may be valuable to design AGI systems with a more rigorous and precise notion of "goal" than humans possess playing a central (though not necessarily dominating) role in their dynamics.

In the context of human intelligence, the concept of a "goal" is a descriptive abstraction. Humans may adopt goals for a time and then drop them, may pursue multiple conflicting goals simultaneously, and may often proceed in an apparently goal-less manner. Sometimes the goal that a person appears to be pursuing, may be very different from the one they think they're pursuing. Evolutionary psychology (Barrett et al. 2002) argues that, directly or indirectly, all humans are ultimately pursuing the goal of maximizing the inclusive fitness of their genes – but given the complex mix of evolution and self-organization in natural history (Salthe 1993), this is hardly a general explanation for human behavior. Ultimately, in the human context, "goal" is best thought of as a frequently useful heuristic concept.

AGI systems, however, may be designed with explicit goal systems. This provides no guarantee that said AGI systems will actually pursue the goals that their goal systems specify – depending on the role that the goal system plays in the overall system dynamics, other dynamic phenomena might sometimes intervene and cause the system to behave in ways opposed to its explicit goals. However, we submit that this design sketch provides a better framework than would exist in an AGI system closely emulating the human brain.

We realize this point may be somewhat contentious – a counter-argument would be that (given that society exists) the human brain is known to support at least moderately ethical behavior, judged by human ethical standards, whereas the ethical propensity of less brain-like AGI systems is not well understood. However, the obvious counter-counterpoints are that:

- Humans are often not consistently ethical, so creating AGI systems potentially much more practically powerful than humans, but with closely human-like ethical, motivational and goal systems, could pose significant risk. People put in positions of gross power imbalance without oversight can often succumb to abusing their power (Zimbardo 2007).

- The effect on a human-like ethical/motivational/goal system of increasing the intelligence, or changing the physical embodiment or cognitive capabilities, of the agent containing the system is difficult to predict, given the complexities involved. Consider a human who could outwit the rest of humanity. Without a social contract to abide by, they may discard ethical behaviour in favor of their personal wants.

The course we tentatively recommend, and are following in our own work, is to develop AGI systems with explicit, hierarchically-modulated goal systems. That is:

- Create one or more "top goals."

- Have the system derive subgoals from these, using its own intelligence, although potentially guided by educational interaction or explicit programming.

- Have a significant percentage of the system's activity governed by the explicit pursuit of these goals.

In addition, these goals should be relatively stable. One way is to represent the goals in the context of a network of related concepts instead of a simplistic representation that requires a quantitative variable (perhaps representing energy available or "hunger") to remain above a threshold.

Included in the "top goals" should be expansion of the conceptual understanding of the other "top goals," as well as understanding the relationship between the goals. An AGI may recognize when goals conflict, and then optimize a balance between them instead of wildly oscillating been fulfillment of two contrary goals.

An important decision regards the immutability of these "top goals." Embedding the goals in a network of concepts will shape their meaning, but conversely will provide resilience against removal. For example, if the network of related concepts describes the goal well enough then they semantically could have a similar implicit influence on the the goal system to that of the goal itself. However, this may be dependent in large part on the architecture of an AGI.

Note that the "significant percentage" in the third point need not be 100 per cent; OpenCog, for example, is designed to combine explicitly goal-directed activity with other "spontaneous" activity. Requiring that all activity be explicitly goal-directed may be too strict a requirement to place on AGI architectures, especially when the route to achieving any particular goal is unclear. Such statements of undirected behavior may set off alarm bells for proponents of provably Friendly AI; however, spontaneous behaviour could still be checked to ensure it isn't predicted to have harmful impacts on the rest of the AGI's goal system.

The next step, of course, is for the top-level goals to be chosen in accordance with the principle of human-Friendliness. The next of our seven points, about the Global Brain, addresses one way of doing this. In our near-term work with OpenCog, we are using somewhat simplistic approaches, with a view toward early-stage system testing.

For instance, an OpenCog agent in a virtual world may have top-level goals to satisfy various physiological demands. At the most basic level, one of these goals usually relates to satisfying an energy demand in order to remain a functional non-entropic system. The complete motivation system currently used for guiding avatars in a virtual sandbox world comes from Dietrich Dörner's Psi theory of motivation and actions (Dörner 2002). Another example of a demand is "certainty," to encourage exploration of the unknown, and increase the extent to which the avatar understands and can predict the world. Other more subtle motivations might be to find more compact representations of patterns observed in the world, so as to minimize the resources required to reason about them.

There are also some motivations that do relate to social behavior, such as the "affiliation" demand that is related to seeking the company and attention of other agents. However, these are not currently chosen specifically to promote friendliness. Demands such as that of maintaining energy clearly need to be tempered with those that encourage socially responsible behaviour, perhaps with relative ordering of their importance. In our current experimental systems, we are primarily concerned with understanding the interplay of these demands and how they influence agent behavior. The extent to which general principals of goal dynamics can be elucidated has yet to be seen.

4.4 Ensure that the early stages of recursive self-improvement occur relatively slowly and with rich human involvement

One of the more exciting, and worrisome, aspects of advanced AGI is the possibility of radical, rapid self-improvement, or "self-programming." If an AGI can understand its own source code and the algorithms underlying its intelligence, it may be able to improve these algorithms and modify its code accordingly, thus increasing its intelligence. Potentially this could lead to a dynamic of accelerating intelligence, in which the smarter the system gets, the better it is at making itself smarter. It's easy to leap from this potential to visions of an AGI ascending from human-level general intelligence to massive transhumanity in days or even minutes. Furthermore, fundamental algorithmic improvements to the AGI system's intelligence could synergize with simpler processes like the AGI marshaling additional hardware for its infrastructure, or more complex options like the AGI inventing new forms of computing hardware, commissioning their construction, and then achieving greater intelligence via porting itself to the new hardware.

At the current stage of AGI development, it's difficult to assess the realism of the more extreme hypothetical forms of "recursive self-improvement." Metaphorical analogues to physical situations like nuclear chain reactions are compelling, but obviously are more poetic than scientific. This sort of computational dynamic has never been experimented with before, and could be susceptible to bottlenecks that aren't currently clear. There also could be modes of acceleration that aren't currently clear – for instance a sufficiently intelligent system might discover new principles of physics or computation enabling intelligence increase beyond what current science could predict. Getting more science fictional, a sufficiently intelligent AGI could potentially figure out how to get in contact with other civilizations, which might then inject it with a large amount of intelligence and knowledge already developed elsewhere.

However, in spite of the particularly speculative nature of this possibility, it seems worthwhile to structure AGI development in such a way as to minimize the associated risks. The primary course one can take in this regard, it seems, is to minimize the extent to which self-improvement and self-programming occur in isolation.[1] The more the AGI system is engaged with human minds and other AGI systems in the course of its self-modification, presumably the less likely it is to veer off in an undesired and unpredictable direction. Of course, this can only happen if we create AGIs that have a strong in-built motivation to consult and work with others in the context of their self-modifications. This may slow down the process of self-improvement, compared to what would maximally be achievable otherwise, but it should also reduce the chance of things rapidly going awry.

A rich coupling between the AGI system and the outside world, such as we've suggested here leads us to our next topic, the Global Brain.

4.5 Tightly link AGI with the Global Brain

Some futurist thinkers, such as Francis Heylighen, believe that engineering AGI systems is at best a peripheral endeavor in the development of novel intelligence on Earth, because the real story is the developing Global Brain (Heylighen 2007; Goertzel 2001): the composite, self-organizing information system comprising humans, computers, data stores, the Internet, mobile phones and other communication systems. Our own views are less dismissive of AGI – we believe that AGI systems will display capabilities fundamentally different from the Global Brain, and that ultimately (unless such development is restricted) self-improving AGI systems will develop intelligence vastly greater than any system possessing humans as a significant component. However, we do respect the power of the Global Brain, and suspect that the early stages of development for an AGI system may go quite differently if it is tightly connected to the Global Brain, via making rich and diverse use of Internet information resources and communication with a diverse section of humanity.

The Internet component of the Global Brain allows us to access almost any piece of information, but we don't have it contextually integrated with our experience, which might be considered the difference between knowing a fact and understanding its meaning. Despite our impressive capacity to generate relationships and perform this integration, we are confined to the neural capacity of our brains and the firing speeds of our neurons. The drop in conceptual bandwidth from brain to keyboard and mouse, or even to human face-to-face communication, reduces our effective ability to reason using knowledge outside of our brain.

Despite the difference in memorization and integrated knowledge, there is great potential for Global Brain integration to bring intelligence enhancement to AGIs. The ability to invoke web searches across documents and databases can greatly enhance the perceived cognitive ability of the AGI, as well as the capability to consult specialized databases like Geographic Information Systems and other specialized services via its web API. Goertzel (2008) reviews the potential for embodied language learning achievable via using AGIs to power non-player characters in widely-accessible virtual worlds or massively multiplayer online games. Work by Orkin and Roy (2010) provides one example. Here, the the actions of interacting humans are used to record social behavior and language, as human players interact with one another within the confines of a restaurant, with the hope that it will provide a contextual grounding to language use.

There is another powerful potential benefit from the Internet for the development of ethical AGI which has two aspects:

1. In a manner analogous to language learning, an AGI system may receive ethical training from a wide variety of humans in parallel, e.g. via controlling characters in wide-access virtual worlds, and gain feedback and guidance regarding the ethics of their behaviors. This could also be weighted by the perceived social acceptance of that character by other human controlled characters. Of course, measuring such a trait is not trivial, but heuristics such as how often others initiate follow-up interactions with a human character, or the frequency that others facilitate the AGI reaching its goals, may be starting points.

2. Internet-based information systems (such as, but not limited to, social media and Wikipedia) may be used to explicitly gather a consensus regarding human values and goals, which may then be appropriately utilized as input for an AGI system's top-level goals.

The second point begins to make abstract-sounding notions like Coherent Extrapolated Volition and Coherent Aggregated Volition, mentioned above, seem more practical and concrete. Gathering information about individuals' values via brain imaging, if such a technology becomes available, is an interesting prospect; but at present, more prosaic methods, such as directly asking people questions, and assessing their ethical reactions to various real-world and hypothetical scenarios, may be used. In addition, engaging people in structured interactions might be used, aimed specifically at eliciting collectively acceptable value systems (see next sub-section). This sort of approach could realize CAV in a practical way, using existing knowledge in social science and psychology.

This resonates with current trends to encourage open governance (Lathrop and Ruma 2010), which suggests that information technology can allow for governments to be more transparent and participatory in their operation. A refinement of this idea is open-source governance, which advocates the application of philosophies from the free software movement to facilitate citizens becoming more active in the legislative process. For example, in 2007, the New Zealand Police invited the public to participate in the drafting of a new Policing Act using online wiki software, before presenting a draft to parliament (BBC 2007). Similar software and community tools to those used to facilitate the involvement of the public with open governance[2] may be useful for gathering value data aimed to assist in shaping AGI goal system content.

Value data can also be gathered in response to an event. Social media allow for real-time commentary on decisions and news from around the world, and analysis that aggregates the general sentiment of groups is becoming more common. Such aggregate analysis is invaluable to marketing departments as it becomes infeasible to comb individually through a deluge of responses to business decisions. Similarly, discourse analysis provides several methods to view in aggregate the composition of phrases and wording used to describe events and topics. While there has been suspicion about the value of sentiment analysis, and whether it reflects a true aggregate sentiment for a population, it seems inevitable that such technology will continue to improve in step with any advancement in AGI.

4.6 Foster deep, consensus-building interactions and commensurability between divergent viewpoints

Three potentially problematic issues arising with the notion of using Global Brain related technologies to form a "coherent volition" from the divergent views of various human beings are:

- The tendency of the Internet to encourage people to interact mainly with others who share their own narrow views and interests, rather than a more diverse body of people with substantially different viewpoints. The three hundred people in the world who want to communicate using predicate logic can find each other,[3] obscure musical virtuosos from around the world can find an audience, and researchers in obscure domains can share papers without needing to wait years for paper journal publication, etc. People will tend to clique with those they share interests with, but eschew people with radically different viewpoints.

- The tendency of many contemporary Internet technologies to reduce interaction to a very simplistic level (e.g. 140-character tweets, brief Facebook wall posts). This allows more topics to be reviewed but can lead to information overload where careful reading is replaced by quick skimming. Such trends mean that a *deep sharing of perspectives* by individuals with divergent views is not necessarily encouraged. As a somewhat extreme example, many of the YouTube pages displaying rock music videos are currently littered with comments by "haters" asserting that rock music is inferior to classical or jazz or whatever their preference is – obviously this is a far cry from deep and productive sharing between people with different tastes and backgrounds. Twitter arguments often end with one or both parties admitting 140 characters is insufficient to convey their views. Fortunately, these sometimes do lead to an agreement to move to email or face-to-face discussion.

- Search engines and content customization algorithms have been criticized as providing overly personalized views of the world. To an extent, we are presented with a viewpoint of the world that makes us comfortable rather than substantially challenging our existing beliefs (Pariser 2011).

Tweets and YouTube comments have their place in the cosmos, but they probably aren't ideal in terms of helping humanity to form a coherent collective volition suitable for providing guidance to an AGI's goal system.

A description of communication at the opposite end of the spectrum is presented in Adam Kahane and Peter Senge's excellent book *Solving Tough Problems* (2004), which describes a methodology that has been used to reconcile deeply conflicting views in some challenging real-world situations (e.g. helping to peacefully end apartheid in South Africa). A core idea of the methodology is to have people with different views explore various possible future scenarios together, in great detail. In cognitive psychology terms, this is a collective generation of hypothetical episodic knowledge. This has multiple benefits, including:

- Emotional bonds and mutual understanding are built in the process of collaboratively exploring the scenarios.
- The focus on concrete situations helps to break through some of the counterproductive abstract ideas that people (on both sides of any dichotomy) may have formed.

- Emergence of conceptual blends that might never have arisen from people with only a single point of view.

The result of such a process, when successful, is not an "average" of the participants' views, but more like a "conceptual blend" of their perspectives.

According to conceptual blending, which some hypothesize to be the core algorithm of creativity (Fauconnier and Turner 2002), new concepts are formed by combining key aspects of existing concepts – but doing so judiciously, carefully choosing which aspects to retain, so as to obtain a high-quality, useful, and interesting new whole. A blend is a compact entity that is similar to each of the entities blended, capturing their "essence" but also possessing its own, novel holistic integrity. In the case of blending different peoples' world views to form something new that everybody is going to have to live with (as in the case of finding a peaceful path beyond apartheid for South Africa, or arriving at a humanity-wide CBV to use to guide an AGI goal system), the trick is that everybody has to agree that enough of the essence of their own view has been captured.

This leads to the question of how to foster deep conceptual blending of diverse and divergent human perspectives, on a global scale. One possible answer is the creation of appropriate Global Brain oriented technologies – but moving away from technologies like Twitter that focus on quick and simple exchanges of small thoughts within affinity groups. On the face of it, it would seem that what's needed is just the opposite – long exchanges about difficult concepts between individuals with radically different perspectives. The exchanges should include sharing feelings, and should take place between people who would not commonly associate with each other.

There is some hope to get there, however. When mailing lists and comments on blogs are carefully moderated to enforce civility between opposing sides of a strongly heated exchange, these can sometimes be a productive way for two conflicted groups to discuss their viewpoints. A good moderator will censor people who resort to insults and detract from the open sharing of arguments, encouraging people to maintain a respectful attitude if they want their arguments to be heard. People have to feel safe sharing ideas, or a true representation of how they feel about something may be self-censored to avoid social attack, or alternatively they'll resort to silence or violence (Patterson 2002). Sometimes people have to agree to disagree in such discussions, but hopefully they still come away having had some of their assumptions questioned, and it becomes clear where the point of divergence in values occurs. If an AGI moderated discussion between groups representing viewpoints of the Global Brain, then perhaps such democratic consensus could more easily be reached.

Building and effectively popularizing Internet technologies that have an ever richer capability to promote this kind of meaningful interaction – and quickly enough to provide guidance to the goal systems of the first highly powerful AGIs – seems a significant, though fascinating, challenge.

4.6.1 Relationship with Coherent Extrapolated Volition

Yudkowsky's CEV concept, mentioned above, has been loosely described by its author as follows:

> *In poetic terms, our coherent extrapolated volition is our wish if we knew more, thought faster, were more the people we wished we were, had grown up farther together; where the extrapolation converges rather than diverges, where our wishes cohere rather than interfere; extrapolated as we wish that extrapolated, interpreted as we wish that interpreted.* (Yudkowsky 2004)

While a moving humanistic vision, this seems to us rather difficult to implement in a computer algorithm in a compellingly "right" way. With many different ways of implementing CEV, the choice between them would involve multiple, highly subtle and non-rigorous human judgment calls.[4] However, if a deep collective process of interactive scenario sharing and analysis is carried out, to arrive at a form of

Coherent Blended Volition, this process may well involve many of the same kinds of extrapolation that are conceived to be part of Coherent Extrapolated Volition. The core difference between the two approaches is that, in the CEV vision, the extrapolation and coherentization are to be done by a highly intelligent, highly specialized software program, whereas in the approach suggested here, these are to be carried out by collective activity of humans as mediated by Global Brain technologies. Our perspective is that the definition of collective human values is probably better carried out via human collaboration, rather than delegated to a machine optimization process; and also that the creation of Internet technologies that can support deep sharing and engagement with humanity, while a difficult task, is significantly easier and more likely to be done in the near future than the creation of narrow AI technology capable of effectively performing CEV style extrapolations.

Another issue with the original CEV formulation is that it assumes that humanity's wishes, when extrapolated, will cohere and converge. There seems no strong basis to believe this will happen – nor to believe that, in the case where it does happen, the extrapolation will reflect human desires or needs in any humanly comprehensible sense. Humans may share ancestral biological goals, but as we discuss above, we have divergent views on many topics. Even people whom we might judge to have an equivalent level of intelligence can differ significantly in their ethics, as our belief systems are built up from a different temporal ordering of experiences, and may be founded on different childhood axioms of social interaction.

4.6.2 Musings on concrete manifestations

Direct democracy is a political system that supports the hands-on involvement of citizens in making decisions on public policy. Full involvement is rarely a good idea though, because very few individuals actually want, or have the time available, to make involved and informed decisions about the minutiae of government policy. AGI-facilitated decisions in such a political environment may instead be based on personal profiles or shadows of our beliefs, but always allowing for the direct involvement of an individual in contentious issues. Indeed, on contentious issues, the AGI may use this as a cue to facilitate deeper discussions between conflicting viewpoints to allow more effective conceptual blending.

Another approach that could be used is the Delphi method, used in forecasting and public-policy making (Linstone and Turoff 2002). This uses a panel of anonymous experts to either forecast or decide recommendations for public policy. These experts provide submissions before having their views integrated and summarized by a facilitator. An AGI could play the role of the facilitator, similar to the earlier suggestion of an AGI being a curator for a mailing list discussion, summarizing views and identifying conflicts to be reviewed in the next round. Due to the assumed vast intelligence of the AGI, the Delphi method could be scaled to a much greater number of experts, including effective selection of participants to choose those who have been identified to have divergent views.

There are many other tools and research methods used in public and social policy that may inspire approaches to building AGIs that can resolve differences in value systems. Frequently policy decisions have to take into account various viewpoints that are at odds, and unlikely to cohere on their own.

4.7 Create a mutually supportive community of AGIs

Omohundro (2009) argues that game-theoretic dynamics related to populations of roughly equally powerful agents, may play a valuable role in mitigating the risks associated with advanced AGI systems. Roughly speaking, if one has a society of AGIs rather than a single AGI, and all the members of the society share roughly similar ethics, then if one AGI starts to go "off the rails," its compatriots will be in a position to correct its behavior.

One may argue that this is actually a hypothesis about which AGI designs are safest, because a "community of AGIs" may be considered a single AGI with an internal community-like design. However, the matter is more subtle than that if one considers the AGI systems embedded in the Global Brain and

human society. In this case, there is some substance to the notion of a population of AGIs systematically presenting themselves to humans and non-AGI software processes as separate entities.

Unfortunately, a society of AGIs is no protection against a single member undergoing a "hard takeoff" and drastically accelerating its intelligence as it simultaneously shifts its ethical principles. In this sort of scenario, with a single AGI that rapidly becomes much more powerful and differently oriented than the others, the latter are left impotent to act so as to preserve their values. This, however, may be mitigated by the difficulty an individual node in a largely homogeneous community may have in trying to subvert the resources still available to the others to mount an immune response or social reprisal. The chance of the community being able to respond to the outlier in time brings up the point considered next, regarding "takeoff speed."

The operation of an AGI society may depend somewhat sensitively on the architectures of the AGI systems in question. Community moderation will work better if the AGIs have a relatively easy way to inspect and comprehend much of the contents of each others' minds. This introduces a bias toward AGIs that rely heavily on more explicit forms of knowledge representation.

The ideal in this regard would be a system like Cyc (Lenat and Guha 1990) with a fully explicit logic-based knowledge representation using a standard ontology – in this case, every Cyc instance would have a relatively easy time understanding the inner thought processes of every other such instance. However, most AGI researchers doubt that fully explicit approaches like this will ever be capable of achieving advanced AGI using feasible computational resources. OpenCog uses a mixed representation, with an explicit (but uncertain and experientially adaptable) logical aspect as well as an explicit subsymbolic aspect more analogous to attractor neural nets.

The OpenCog design also includes a yet unimplemented mechanism called *Psynese*, intended to make it easier for one OpenCog instance to translate its personal thoughts into the mental language of another. This translation process may be quite subtle, since each instance will generally learn a host of new concepts based on its experience, and these concepts may not possess any compact mapping into shared linguistic symbols or percepts. The wide deployment of some mechanism of this nature, among a community of AGIs, will be very helpful in enabling this community to display the level of mutual understanding needed for strongly encouraging ethical stability.

Of course, it is possible that the distinction between an individual instance and a community will not be meaningful. OpenCog is currently a mostly non-distributed application,[5] but there is a lot of interest in moving it to a distributed architecture that can take advantage of cloud services or volunteers willing to run local instances of OpenCog in a manner similar to SETI@home or other distributed processing tasks.

Here there may be even more concern about unscrupulous individuals injecting knowledge intended to compromise the AGI's non-partisan friendliness; however, there are a number of similar areas that are eager to ensure consistent behavior in a networked application. BitTorrent is a heterogeneous network of peer-to-peer file sharing clients. While the core protocol between clients is shared, individual clients have a number of guards to prevent "leechers" who don't contribute to the health of the torrents and the performance of the network (Legout et al. 2007), and the distributed currency, Bitcoin, uses cryptography techniques to ensure the safe transaction of currency without a centralized bank (Nakamoto 2008). A key point is that the health of these networks relies on the majority of its nodes acting in a "friendly" way, consistent with the original design. For a massively distributed and large network, the chances of any one individual being able to co-opt enough networking elements or computing power become exceedingly small, although not impossible.

4.8 Encourage measured co-advancement of AGI software and AGI ethics theory

Our next point intersects with all the previous ideas. Everything involving AGI and Friendly AI (considered together or separately) currently involves significant uncertainty, and it seems likely that significant revision of current concepts will be necessary as progress on the path toward powerful AGI proceeds. However, whether there is time for such revision to occur, before AGI at the human level or above is created, depends on how fast the progress toward AGI turns out to be. Ideally, progress would be slow enough that, at each stage of intelligence advance, concepts such as those discussed in this paper can be re-evaluated and re-analyzed in the light of the data gathered, and AGI designs and approaches can be revised accordingly as necessary.

But how can this kind of measured co-advancement be encouraged? Of course, an all-powerful world government could strictly enforce such a pattern of development, but that's far from the current situation, and history shows that such scenarios tend to have their downsides.

One possibility would be the creation of an "AGI Nanny" – i.e. an AGI system with at least human level intelligence, and with a high level of power to surveil and police the human world, but no taste for radical self-modification or reproduction. Such an AGI Nanny would allow human science, technology, and thought to advance fairly freely. It would, however, have the goal of forcibly preventing any development that seemed to be leading to premature creation of highly advanced AGI or other Singularity-enabling technologies, such as molecular nanotechnology or radical human brain enhancement. Once a sufficient understanding of AGI architecture and ethics was achieved, the AGI Nanny would release its control and enable the Singularity to proceed. The "Nanny" metaphor is chosen specifically because a human nanny watches human children until they grow up, then it releases them. Similarly, the AGI Nanny would watch over the human species until its technological and conceptual framework matured sufficiently to enable it to launch a Singularity with acceptable safety.

There seems no in-principle reason why an AGI Nanny scenario like this couldn't succeed. An AGI system like OpenCog, with an explicit goal system guiding most of its behaviors, could be programmed with goals consistent with the AGI Nanny role. However, the idea obviously would face significant practical obstacles (how would the AGI Nanny come to power, in an ethical way?), and also has significant risk attached to it (what if, in spite of our best efforts, the AGI Nanny behaved in some unpredictable way?).

Of course, to make an AGI Nanny would itself require dramatic advances in AGI technology beyond the current state of the art. It's unclear whether it's easier to create an AGI Nanny than to create an unpredictably and dramatically self-improving AGI system. It may be that a similar set of technologies (OpenCog for instance) could be used to create either one, leading to possible scenarios where an AGI Nanny focused team competes with a team focused on more hurriedly launching a Singularity via a rapidly self-improving AGI, each team using a separate forked version of the same codebase.

Setting aside the AGI Nanny possibility, how might measured co-advancement of AGI technology and AGI ethics understanding best be encouraged?

4.9 Develop advanced AGI sooner not later

Somewhat ironically, it seems that one good way to ensure that AGI development proceeds at a relatively measured pace may be to initiate serious AGI development sooner rather than later. The same AGI concepts will yield slower practical development today than ten years from now, and be slower ten years from now than twenty years from now. This is a result of the ongoing rapid advancement of various tools related to AGI development, such as computer hardware, programming languages, and computer science algorithms.

Currently, the pace of AGI progress is sufficiently slow that practical work towards human-level AGI is in no danger of outpacing associated ethical theorizing. However, if we want to avoid the future occurrence of this sort of dangerous outpacing, our best practical choice is to ensure more substantial AGI development occurs in the phase before the development of tools and hardware that will make AGI development and prototyping much quicker. Which, of course, is why the authors are doing their best in this direction via their work on the OpenCog project.

Furthermore, this point is connected with the need, raised above, to foster the development of Global Brain technologies that "foster deep, consensus-building interactions between divergent viewpoints." For this sort of technology to be maximally valuable, it is necessary that it be created quickly enough to incorporate the blended volition it extracts, so that we can use it to shape the goal system content of the first powerful AGIs. In essence, we want both deep-sharing Global Brain technology and AGI technology to evolve together rapidly, in comparison to the ongoing improvement in computing hardware and software engineering tools. Such a goal is challenging, since the latter aspects are more easily incrementally improved and thus receive dramatically more funding and focus than AGI.

5. Conclusion

We have briefly considered a number of strategies oriented toward increasing the odds of AGI Friendliness, with a focus on techniques relevant to an open-source AGI project such as OpenCog. None of these strategies gives any guarantees, but combined they should bias the odds in favor of a positive outcome.

None of the nine points raised is particularly well understood, but research into each of them could proceed meaningfully, in large part separately from any particular AGI design. However, little attention seems to get paid to AGI-ethics issues at present, either by the research funding establishment or by individual researchers on their own time.

The Singularity Institute for AI (http://singinst.org), co-founded by Yudkowsky who was referenced frequently above, is doing an admirable job of pursuing research in the Friendly AI domain. However, as its perspective is significantly different from the one described here, much of its work does not yet directly address the issues raised above. Although SIAI initially co-founded the OpenCog project, recently its preference is for closed approaches to AGI development. SIAI's recent research foci have included CEV, the pursuit of provably Friendly AGI designs, broad exploration of existential risks, and advocacy of general human rationality. While worthy, each of these covers only a small scope of the area required to figure out how to bias an open-source (or otherwise) AGI software project in the direction of Friendliness.

We anticipate that once AGI research advances to the point where it demonstrates more exciting practical capabilities, and more resources are consequently focused on AGI, then more resources will also be focused on problems related to AGI-ethics. We have proposed that the best approach is to begin the serious co-evolution of functional AGI systems and AGI-related ethical theory, as well as the development of deep-sharing-oriented Internet tools, as soon as possible. These need to move ahead before we have so much technical infrastructure and computing power that parties relatively unconcerned with ethics are able to rush forward with AGI development in dangerously hasty fashion.

Notes

1. Ironically, this is almost the opposite of one approach sometimes suggested for regulating AGI development: AI boxing. This is where an AGI is confined to a tightly controlled environment with limited interaction with the rest of the world.

2. For a selection of these, see the Metagovernment wiki, available at http:\\metagovernment.org (accessed December 26, 2011).

3. See the website Lobjan: The Logical Language, available at http://lojban.org (accessed December 26, 2011).

4. Readers are encouraged to look at the original CEV essay (Yudkowsky 2004) online and make their own assessments.

5. Various narrow AI components, or parts of the embodiment architecture, can run as separate servers, but the core control resides in a single application.

References (all online sources last accessed December 26, 2011)

Adams, Sam, Itamar Arel, Joscha Bach, Robert Coop, Rod Furlan, Ben Goertzel, J. Storrs Hall, Alexei Samsonovich, Matthias Scheutz, Matt Schlesinger, Stuart C. Shapiro, and John Sowa. 2012. Mapping the landscape of artificial general intelligence. *AI Magazine*. Forthcoming.

Arkin, Ronald. 2009. *Governing lethal behavior in autonomous robots*. London: Chapman and Hall/CRC.

Barrett, Louise, Robin Dunbar, and John Lycett. 1993. *Human evolutionary psychology*. New Haven: Princeton University Press.

BBC. 2007. NZ police let public write laws. BBC News 26 September.
Available at http://news.bbc.co.uk/2/hi/7015024.stm.

Bostrom, Nick. Existential risks. 2002. *Journal of Evolution and Technology* 9(1) (March).
Available at http://www.jetpress.org/volume9/risks.html.

De Garis, Hugo. 2005. *The artilect war: cosmists vs. terrans : a bitter controversy concerning whether humanity should build godlike massively intelligent machines*. Palm Springs, CA: ETC Publications.

Drexler, K. E. 1986. *Engines of creation: The coming era of nanotechnology*. New York: Anchor Books.
Available at http://e-drexler.com/p/06/00/EOC_Cover.html.

Dörner**,** Dietrich. 2002. *Die Mechanik des Seelenwagens. Eine neuronale Theorie der Handlungsregulation*. Bern: Verlag Hans Huber.

Fauconnier, Gilles and Mark Turner. 2002. *The way we think: Conceptual blending and the mind's hidden complexities*. New York: Basic Books.

Gilligan, Carol. 1982. *In a different voice*. Cambridge, MA: Harvard University Press.

Goertzel, Ben. 2001. *Creating internet intelligence*. New York: Plenum Press.

—. 2006. *The hidden pattern*. Boca Raton, FL: BrownWalker Press.

—. 2008. A pragmatic path toward endowing virtually-embodied AIs with human-level linguistic capability. IEEE World Congress on Computational Intelligence (WCCI).

—. 2009a. The embodied communication prior. In *Proceedings of ICCI-09*. Hong Kong.

—. 2009b. AGI, ethics, cognitive synergy and ethical synergy. The multiverse according to Ben, comment posted September 9.
Available at http://multiverseaccordingtoben.blogspot.com/2009/09/agi-ethics-cognitive-synergy-and.html.

—. 2010. Coherent aggregated volition. The multiverse according to Ben, comment posted 12 March.
Available at http://multiverseaccordingtoben.blogspot.com/2010/03/coherent-aggregated-volition-toward.html.

Goertzel, Ben, and Stephan Vladimir Bugaj. 2008. Stages of ethical development in uncertain inference based AI systems. *In Proceedings of First AGI Conference*. Amsterdam: IOS Press: 448-59.

Ben Goertzel, Hugo de Garis, Cassio Pennachin, Nil Geisweiller, Samir Araujo, Joel Pitt, Shuo Chen, Ruiting Lian, Min Jiang , Ye Yang, and Deheng Huang. 2010. OpenCogBot: Achieving generally intelligent virtual agent control and humanoid robotics via cognitive synergy. *Proceedings of ICAI 2010*, Beijing.

Hart, David, and Ben Goertzel. 2008. Opencog: A software framework for integrative artificial general intelligence. In AGI, volume 171 of *Frontiers in Artificial Intelligence and applications*. Amsterdam: IOS Press, 2008: 468-72.
Available at http://dblp.uni-trier.de/db/conf/agi/agi2008.html#HartG08.

F. Heylighen. 2007. The global superorganism: An evolutionary-cybernetic model of the emerging network society. *Social Evolution and History* 6(1): 57-117.

Hibbard, Bill. 2002. *Superintelligent machines*. New York: Kluwer Academic.

Kohlberg, Lawrence. 1981. *The philosophy of moral development: Moral stages and the idea of justice*. New York: Harper and Row.

Laird, John, Robert Wray, Robert Marinier, and Pat Langley. 2009. Claims and challenges in evaluating human-level intelligent systems. *Proceedings of AGI-09*.

Lathrop, Daniel and Laurel Ruma, ed. 2010. *Open government: Transparency, collaboration and participation in practice*. Sebastopol, CA: O'Reilly Media.

Legg, Shane. 2006a. Unprovability of friendly AI. Vetta Project, comment posted September 15.
Available at http://www.vetta.org/2006/09/unprovability-of-friendly-ai/.

—. 2006b. Friendly AI is bunk. Vetta Project, comment posted September 9.
Available at http://commonsenseatheism.com/wp-content/uploads/2011/02/Legg-Friendly-AI-is-bunk.pdf.

Legout, Arnaud, Nikitas Liogkas, Eddie Kohler, and Lixia Zhang. 2007. Clustering and sharing incentives in BitTorrent systems. In *Sygmetrics '07: Proceedings of the 2007 ACM Sigmetrics International Conference on Measurement and Modeling of Computer Systems*: 301-312. New York: ACI.

Lenat, Douglas, and R. V. Guha. 1990. *Building large knowledge-based systems: Representation and inference in the Cyc Project*. Reading, PA: Addison-Wesley.

Linstone, Harold A., and Murray Turo. 2002. The Delphi method: Techniques and applications. Available at http://is.njit.edu/pubs/delphibook/.

Nakamoto, Satoshi. 2008. Bitcoin: A peer-to-peer electronic cash system. Bitcoin.org. Available at http://www.bitcoin.org/bitcoin.pdf.

Omohundro, Stephen. 2008. The basic AI drives. *Proceedings of the First AGI Conference*, edited by Ben Goertzel and Pei Wang. Amsterdam: IOS Press.

—. 2009. Creating a cooperative future. Self-aware systems: Understanding natural and artificial intelligence, comment posted February 23. Available at http://selfawaresystems.com/2009/02/23/talk-on-creating-a-cooperative-future/.

Orkin, Jeff and Deb Roy. 2010. The restaurant game: Learning social behavior and language from thousands of players online. *Journal of Game Development* 3(2): 39-60.

Pariser, Eli. 2011. *The filter bubble: What the internet is hiding from you*. New York: Viking.

Patterson, Kerry, Joseph Grenny, Ron McMillan, and Al Switzler. 2002. *Crucial conversations: Tools for talking when stakes are high*. New York: McGraw-Hill.

Raymond, Eric Steven. 2000. The cathedral and the bazaar, version 3.0. Thyrsus Enterprises. Available at http://www.catb.org/~esr/writings/homesteading/cathedral-bazaar/.

Rizzolatti, Giacomo, and Laila Craighero. 2004. The mirror-neuron system. *Annual Review of Neuroscience* 27: 169-92.

Salthe, Stanley N. 1993. *Development and evolution: Complexity and change in biology*. Cambridge, MA: MIT Press.

Sotala, Kaj. 2011. 14 objections against AI/friendly AI/the Singularity answered. Xuepolis. Available at http://www.xuenay.net/objections.html.

Wallach, Wendell, and Colin Allen. 2010. *Moral machines: Teaching robots right from wrong*. New York: Oxford University Press.

Waser, Mark R. 2008. Discovering the foundations of a universal system of ethics as a road to safe artificial intelligence. *Proceedings of BICA 2008*. Arlington, VA. Available at http://becominggaia.files.wordpress.com/2010//06/waser-bica08.pdf.

Yudkowsky, Eliezer. 2001. Creating Friendly AI 1.0: The analysis and design of friendly goal architectures. Singularity Institute for Artificial Intelligence. Available at http://singinst.org/upload/CFAI.html.

—. 2004. Coherent extrapolated volition. Singularity Institute for Artificial Intelligence. Available at http://singinst.org/upload/CEV.html.

—. 2006. What is Friendly AI? Singularity Institute for Artificial Intelligence. Available at http://singinst.org/ourresearch/publications/what-is-friendly-ai.html.

Zimbardo, Philip. G. 2007. *The Lucifer effect: Understanding how good people turn evil*. New York: Random House.

Appendix: Tables 1, 2, and 3.

Stage	Characteristics
Pre-ethical	• Piagetian infantile to early concrete (aka pre-operational). • Radical selfishness or selflessness may, but do not necessarily, occur. • No coherent, consistent pattern of consideration for the rights, intentions or feelings of others. • Empathy is generally present, but erratically.
Conventional Ethics	• Concrete cognitive basis. • Perry's Dualist and Multiple stages. • The common sense of the Golden Rule is appreciated, with cultural conventions for abstracting principles from behaviors. • One's own ethical behavior is explicitly compared to that of others. • Development of a functional, though limited, theory of mind. • Ability to intuitively conceive of notions of fairness and rights. • Appreciation of the concept of law and order, which may sometimes manifest itself as systematic obedience or systematic disobedience. • Empathy is more consistently present, especially with others who are directly similar to oneself or in situations similar to those one has directly experienced. • Degrees of selflessness or selfishness develop based on ethical groundings and social interactions.

Table 1: Integrative Model of the Stages of Ethical Development, part 1

Stage	Characteristics
Mature Ethics	• Formal cognitive basis. • Perry's Relativist and "Constructed Knowledge" stages. • The abstraction involved with applying the Golden Rule in practice is more fully understood and manipulated, leading to limited but nonzero deployment of the Categorical Imperative. • Attention is paid to shaping one's ethical principles into a coherent logical system. • Rationalized, moderated selfishness or selflessness. • Empathy is extended, using reason, to individuals and situations not directly matching one's own experience. • Theory of mind is extended, using reason, to counterintuitive or experientially unfamiliar situations. • Reason is used to control the impact of empathy on behavior (i.e. rational judgments are made regarding when to listen to empathy and when not to). • Rational experimentation and correction of theoretical models of ethical behavior, and reconciliation with observed behavior during interaction with others. • Conflict between pragmatism of social contract orientation and idealism of universal ethical principles. • Understanding of ethical quandaries and nuances develop (pragmatist modality), or are rejected (idealist modality). • Pragmatically critical social citizen. Attempts to maintain a balanced social outlook. Considers the common good, including oneself as part of the commons, and acts in what seems to be the most beneficial and practical manner.

Table 2: Integrative Model of the Stages of Ethical Development, part 2

Stage	Characteristics
Enlightened Ethics	• Reflexive cognitive basis. • Permeation of the categorical imperative and the quest for coherence through inner as well as outer life. • Experientially grounded and logically supported rejection of the illusion of moral certainty in favor of a case-specific analytical and empathetic approach that embraces the uncertainty of real social life. • Deep understanding of the illusory and biased nature of the individual self, leading to humility regarding one's own ethical intuitions and prescriptions. • Openness to modifying one's deepest, ethical (and other) beliefs based on experience, reason and/or empathic communion with others. • Adaptive, insightful approach to civil disobedience, considering laws and social customs in a broader ethical and pragmatic context. • Broad compassion for and empathy with all sentient beings. • A recognition of inability to operate at this level at all times in all things, and a vigilance about self-monitoring for regressive behavior.

Table 3: Integrative Model of the Stages of Ethical Development, part 3