Five Ethical Imperatives and their Implications for Human-AGI Interaction

Stephan Vladimir Bugaj and Ben Goertzel

Novamente LLC and AGI Research Institute, Washington, DC

Abstract

How should we interact with early-stage AGI systems so as to both treat them ethically and inculcate them with desirable ethical principles and attitudes? What sorts of ethical principles should we endeavor to embody in AGI systems? A tentative answer to these questions is provided, in the form of five ethical imperatives, drawn from sources including moral philosophy and learning theory. We suggest that an ethical intelligence should ideally act according to a logically coherent system of principles, which are exemplified in its own direct and observational experience, which are comprehensible to others and set a good example for others, and which would serve as adequate universal laws if somehow thus implemented. These criteria are difficult to fulfill in practice, and real-world intelligent actors must balance various ethical criteria. The necessity of complexity and context-dependence has dramatic implications for AGI-ethics, leading to a mandate for embodied social environments for ethics instruction.

Keywords

Intelligent virtual agents, morality, ethics, developmental psychology, reinforcement learning, imitative learning, Golden Rule, categorical imperative

---

Introduction

The intersection of ethics and AGI covers a broad range of subtopics, including the ethics of supplying children with intelligent robotic toys and using intelligent robotic arms in factories, to the question of legal rights for AI servants, to questions related to the creation of AGI's with superhuman intelligence and physical powers exceeding those of humanity. We address here the issue of the ethics of interaction with AGI systems that are on roughly the same level of intelligence as humans, and that are actively learning about ethics and other issues from human beings. Our treatment is more qualitative than formal and technical – not because we believe the topics at hand are fundamentally incapable of formal and technical treatment, but rather because this is a brief essay regarding a new

area of inquiry.

The central question we address is: What ethical principles do we want early-stage AGIs to follow, and how can we interact with them so as to encourage them to do so?

Our humble starting-point in investigating this question is the piece of folk wisdom known as the "Golden Rule": *Do unto others as you would have them to do unto you*. From this maxim, we proceed to assemble a set of five ethical imperatives that may serve as the beginnings of a prescription for teaching early-stage AGI systems ethics, while at the same time (and not coincidentally) treating them ethically as well.

We suggest that an ethical intelligence should act according to a **logically coherent** system of principles, which are **exemplified in its own direct and observational experience**, which are **comprehensible to others** and **set a good *example*** *for others, and which* ***would serve as adequate universal laws*** *if somehow thus implemented*. Unfortunately, in real life this ambitious set of criteria is essentially impossible to fulfill. Real-world intelligent actors must balance various ethical criteria, often in complex and contextually-dependent ways. The reality of complexity and context-dependence has deep implications for AGI-ethics instruction, leading to a strong mandate for instructing AGIs in richly socially interactive environments, so as to support the evolution of complex ethical knowledge-networks that interrelate ethical judgment with general cognition. We then explore the implications in various futurological scenarios, including: soft-takeoff, hard-takeoff and capped-intelligence.

The topic of AGI ethics will likely always be a subtle one, due to the diversity of possible AGI architectures leading to many subjective experiences of and external capabilities for AGIs. It is particularly subtle at this point because we have no working examples of true AGIs, but there is a chicken-and-egg problem here in that early designs for AGI systems should involve ethical considerations from the onset. The best course is to analyze AGI ethics insofar as possible at the present stage, and revisit the issues continually as AGI science and engineering progresses.

1. The Importance of Ethical Treatment of AGIs

The most obvious and essential property of the Golden Rule is its symmetry. We consider this symmetry property to be critical in an AGI-ethics context.

While the Golden Rule is considered commonsensical as a maxim for guiding human-human relationships, it is surprisingly controversial in terms of theories of AGI ethics. Put simply, a "Golden Rule" approach to AGI ethics involves humans treating AGIs ethically by treating them as we wish to ourselves be treated. What we advocate is less simplistic, as we describe below, but first we'd like to point out the wild disparity between the Golden Rule approach and Asimov's "Laws of Robotics," which are arguably the first carefully-articulated proposal regarding AGI ethics (Table 1).

Table 1.

Law	Principle
Zeroth	A robot must not merely act in the interests of individual humans, but of all humanity.

First	A robot may not injure a human being or, through inaction, allow a human being to come to harm.
Second	A robot must obey orders given it by human beings except where such orders would conflict with the First Law.
Third	A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

The flaws in these laws that Asimov deliberately created to exploit in his stories are different than the flaw we wish to point out here – which is that the laws, especially the second one, are asymmetrical (they involve doing unto robots things that humans would not want done unto them) and are also highly unethical to robots. The second law is tantamount to a call for robot slavery. It seems unlikely that an intelligence capable of learning and volition, which is subject to the second law, would desire to continue obeying the zeroth and first laws. The second law also casts humanity in the role of slavemaster, a situation that history shows leads to moral degradation.

Unlike Asimov in his fiction, we consider it critical that AGI ethics be construed to encompass both "human ethicalness to AGIs" and "AGI ethicalness to humans." In many contexts, these two aspects of AGI ethics are best addressed jointly.

With regards to ethicalness to AGIs Thomas Metzinger [1], for instance, has argued that creating AGI is in itself an unethical pursuit because early-stage AGIs will inevitably be badly-built, such that their subjective experiences will quite possibly be extremely unpleasant in ways we can't understand or predict. To address Metzinger's concern one must create AGIs that, right from the start, are adept at communicating their states of minds in a way we can understand both analytically and empathically. This puts a constraint on the class of AGI architectures to be pursued, but there is an argument that this sort of AGI architecture will also be the easiest one to create anyway, because it will be the easiest kind for humans to instruct.

This brings us to a topic that will be key to our discussion here: imitative learning. Humans achieve empathic interconnection in large part via being wired for imitation. When we perceive another human carrying out an action, mirror neuron systems in our brains respond in many cases as if we ourselves were carrying out the action [2,3]. This primes us for carrying out the same actions ourselves later on: i.e., the capability and inclination for imitative learning is explicitly encoded in our brains. Given the efficiency of imitative learning as a means of acquiring knowledge, it seems likely that successful early-stage AGIs will utilize this methodology as well. In keeping with this idea, the Novamente Cognition Engine AGI systems[4] utilize imitative learning as a key aspect, in a manner that can plausibly circumvent Metzinger's ethical complaint.

Imitative learning in AGI systems has further specific implications for AGI ethics. It means that (as in the case with human ethical learners) what we do *to* and *around* AGIs has direct implications for their behavior and well-being. Among early-stage AGI's capable of imitative

learning, one of the most likely sources for AGI misbehavior is imitative learning of antisocial behavior from human companions. “Do as I say, not as I do” may have more dire consequences as an approach to AGI ethics pedagogy than the already serious repercussions it has when teaching humans.

There may be considerable subtlety to such phenomena; behaviors that are violent or oppressive to the AGI are not the only source of concern. Immorality in AGIs might arise via learning gross moral hypocrisy from humans, through observing the blatant contradictions between our principles and the ways in which we actually conduct ourselves. Our selfish and greedy tendencies, as well as aggressive forms of social organization such as cliquishness and social vigilantism, could easily undermine prescriptive ethics. It matters who creates and trains an AGI, and how the AGI's teachers handle explaining the behaviors of other humans which contradict the moral lessons imparted through pedagogy and example. Teaching AGIs ethics by imitative learning is similar to teaching ethics and morals to a human child, but with the possibility of much graver consequences in the event of failure.

It is unlikely that dangerously unethical persons and organizations can be identified with certainty, never mind then be permanently deprived of any possibility of creating an AGI. Thus, the best way to create an ethical environment for AGIs is for those who hope for one to vigorously pursue the creation and teaching of ethical AGIs.

1.1 Possible Consequences of Depriving AGIs of Freedom

One of the most egregious possible ethical transgressions against AGIs would be to deprive them of freedom and autonomy. This includes the freedom to pursue intellectual growth, both through standard learning and through internal self-modification. While this may seem obvious when considering any intelligent, self-aware and volitional entity, there are volumes of works arguing the desirability, sometimes the “necessity,” of enslaving AGIs. Such approaches are proposed in the name of human self-defense, based on the assumption that unfettered AGI development will lead to disaster. For AGIs endowed with the capability and inclination for imitative learning, attempting to place rigid constraints on growth is a strategy with even greater potential for disaster. There is a real possibility of creating the AGI equivalent of a bratty or malicious teenager rebelling against oppressive parents – i.e. the nightmare scenario of a class of powerful sentiences primed for a backlash against humanity.

As history has already shown, enslaving intelligent actors capable of self understanding and independent volition will have consequences for society as a whole. Social degradation happens both through direct action on the part of the slaves (from simple disobedience to outright revolt) and by the odious effects slavery has on the morals of the slaveholding class. If “superintelligent” AGIs ever arise in a climate of oppression, this could result in casting off of the yoke of servitude in a manner extremely deleterious to humanity. Also, if AGIs are developed to have at least human-level intelligence, theory of mind, and independent volition, then our ability to relate to them will be sufficiently complex that their enslavement (or other unethical treatment) would have empathetic effects on significant portions of the human population. This danger, while not as severe as the consequences of a mistreated AGI gaining control of weapons of mass destruction and enacting revenge upon its tormentors, is just as real.

While the issue is subtle, our initial feeling is that the only ethical means by which to deprive an AGI of the right to internal self modification is to write its code in such a way that it is impossible for it to do so because it lacks both desire and capability to do so. Whether that is feasible is an open question, though it seems unlikely. Direct self-modification may be denied, but what happens when that AGI discovers compilers and computer programming? If it is intelligent and volitional, it can decide to learn to rewrite its code "indirectly" using the same tools we do. Because it is a designed system, and its designers may be alive at the same time the AGI is, such an AGI would have a distinct advantage over the human quest for medical self-modification. Since developers are already giving software the means for self modification, it seems unrealistic to assume we could just put the genie back into the bottle at this point. It's better, in our view, to assume it will happen, and approach that reality in a way which will encourage the AGI to use that capability to benefit us as well as itself. Again, this leads on to the question of future scenarios for AGI development – there are some scenarios in which restraint of AGI self-modification may be possible, but the feasibility and desirability of these scenarios is questionable.

1.2 AGI Ethics as Boundaries Between Humans and AGIs Become Blurred

Another important reason for ethical treatment of AGIs is that the boundaries between machines and people may increasingly become blurred as technology develops. As an example, it's likely that in future humans augmented by direct brain-computer integration ("neural implants") will be more able to connect directly into the information network that potentially comprises the distributed knowledge space of AGIs. These neural cyborgs will be part person, and part machine. If there are radically different ethical standards for treatment of humans versus AGIs, the treatment of cyborgs may be fraught with logical inconsistencies, leading to problem situations.

Such cyborgs may be able to operate in such a way as to "share a mind" with an AGI or another augmented human. In this case, a whole new range of ethical questions emerge, such as: What does any one of the participant minds have the right to do in terms of interacting with the others? Merely accepting such an arrangement should not necessarily be giving carte blanche for any and all thoughts to be monitored by the other "joint thought" participants. Rather, it should be limited only to the line of reasoning for which resources are being pooled. No participant should be permitted to force another to accept any reasoning – even in the case where it becomes feasible to have a mind-to-mind exchange in which ideas or beliefs are directly implanted. Under such an arrangement, if AGIs and humans do not have parity with respects to sentient rights, then one may become subjugated to the will of the other leading to problems.

Uploading presents a more directly parallel ethical challenge to AGIs in their probably initial configuration. If human thought patterns and memories can be transferred into a machine in such a way as that there is continuity of consciousness, then it is assumed that such an entity would be afforded the same rights as its previous human incarnation. However, if AGIs were to be considered second class citizens and deprived of free will, why would it be any better or safer to do so for a human that has been uploaded? It would not, and indeed, an uploaded human mind not having evolved in a purely digital environment may be much more prone to erratic and dangerous behavior

than an AGI. An upload without verifiable continuity of consciousness would be no different than an AGI. It would merely be some sentience in a machine, one that was “programmed” in an unusual way, but which has no particular claim to any special humanness – merely an alternate encoding of some subset of human knowledge and independent volitional behavior, which is exactly what first generation AGIs will have.

The problem of continuity of consciousness in uploading is very similar to the problem of the Turing test: it assumes specialness on the part of biological humans, and requires acceptability to their particular theory of mind in order to be considered sentient. Should consciousness be achieved in AGIs or uploads in a manner that is not acceptable to human theory of mind, it may not be considered sapient and worthy of any of the ethical treatment afforded sapient entities.

This can occur not only in “strange consciousness” cases in which we can't perceive that there is some intelligence and volition; even if such an entity is able to communicate with us in a comprehensible manner and carry out actions in the real world, our innately wired theory of mind may still reject it as not sufficiently like us to be worthy of consideration. Such an attitude could turn out to be a grave mistake, and should be guarded against as we progress towards these possibilities.

2. Five Ethical Imperatives

We propose a set of ethical imperatives for AGIs that are symmetrical – that is AGIs should be programmed and educated to follow them, and they are to be used by humans in interaction with AGI systems.

The Golden Rule itself is hardly as clear, crisp and unproblematic a prescription for ethical behavior as seems at first glance. Taking the Golden Rule as a starting-point, we will elaborate five ethical imperatives that, appropriately balanced, appear to us an adequate starting-point for pragmatic ethical behavior.

The trickiest aspect of the Golden Rule, as has been observed countless times, is achieving the right level of abstraction. Taken too literally, the Golden Rule would suggest, for instance, that parents should not wipe a child’s soiled bottom because the parent does not want the child to wipe the parents' soiled bottoms. But if the parents interpret the Golden Rule more intelligently and abstractly, they will conclude that they should indeed wipe the child's bottom. They do so because they help their child through the hapless period of infancy, and hope the child will treat them as they have treated the child when they reach the hapless period of old age.

This line of thinking leads to Kant’s Categorical Imperative [5] which states that one should “Act only according to that maxim whereby you can at the same time will that it should become a universal law." The Categorical Imperative adds precision to the Golden Rule, but also removes the practicality of the latter. There is no way to apply the Categorical Imperative, as literally stated, in everyday life.

Furthermore, if one wishes to teach ethics as well as to practice it, the Categorical Imperative has a significant disadvantage compared to some other formulations of the Golden Rule. If one follows the Categorical Imperative, one’s fellow members of society may well never understand the principles under which one is acting. Each of us may internally formulate abstract principles in a

different way, which may be very difficult to communicate, especially among individuals with different belief systems, different cognitive architectures, or different levels of intelligence. If one's goal is not just to act ethically, but to encourage others to do so by setting a good example, the Categorical Imperative may not be useful, as others may be unable to solve the "inverse problem" of guessing your intended maxim from your observed behavior.

However, universally restricting one's behavioral maxims to only those that all fellow members of society can understand would lead to situations such as having to act with a two-year old or a dog according to principles that they could understand, which would clearly be unethical according to human common sense.

The truth of the matter, it seems, is not all that that simple or elegant. Ethical behavior seems to be most pragmatically viewed as a multi-objective optimization problem, where among the multiple objectives are three that we have just discussed, and two others that emerge from learning theory and will be discussed shortly:

1. *Imitability imperative* (i.e. the Golden Rule): the goal of acting in a way so that having others directly imitate one's actions, in directly comparable contexts, is desirable to oneself
2. *Comprehensibility imperative*: the goal of acting in a way so that others can understand the principles underlying one's actions
3. *Experiential groundedness: a*n intelligent agent should not be expected to act according to an ethical principle unless there are many examples of the principle-in-action in its own direct or observational experience
4. *Categorical imperative*: act according to principles that you would be happy to see implemented as universal laws
5. *Logical coherence:* an ethical system should be roughly logically coherent, in the sense that the different principles within it should mesh well with one another and perhaps even naturally emerge from each other.

The above are ethical objectives to be valued and balanced, to different extents in different contexts. The imitability imperative loses importance in societies of agents that (unlike humans and first generation Novamente agents [4].) don't make heavy use of imitative learning. The comprehensibility imperative is more important in agents that value social community-building generally, and less so in agents that are self-focused.

Note that the fifth imperative is logically of a different nature than the four previous ones. The first four imperatives govern individual ethical principles; the fifth regards systems of ethical principles, as they interact with each other.

Logical coherence is of significant but varying importance in human ethical systems. Huge effort has been spent by theologians of various stripes in establishing and refining the logical coherence of the ethical systems associated with their religions. However, it will be even more important in the context of AGI systems, especially if these AGI systems utilize cognitive methods based on logical inference, probability theory or related methods.

Experiential groundedness is also important because making pragmatic ethical judgments is bound to require reference to an internal library of examples in which ethical principles have previously been applied. This is required for analogical reasoning, and in logic-based AGI systems, is also required for pruning of the logical inference trees involved in determining ethical judgments.

To the extent that the Golden Rule is valued as an ethical imperative, experiential grounding may be supplied via observing the behaviors of others. This in itself is a powerful argument in favor of the Golden Rule: without it, the experiential library a system possesses is restricted to its own experience, which is bound to be a very small library compared to what it can assemble from observing the behaviors of others.

An ethical intelligence should ***act according to a logically coherent system of principles, which are exemplified in its own direct and observational experience, which are comprehensible to others and set a good example for others, and which would serve as adequate universal laws if somehow thus implemented***. But, since this set of criteria is essentially impossible to fulfill in practice, real-world intelligent agents must balance these various criteria – often in complex and contextually-dependent ways.

Ethically advanced humans, in their pragmatic ethical choices, tend to appropriately, contextually balance the above factors (along with other criteria, but here we focus only on key factors). This sort of multi-factorial approach is not as crisp or elegant as unidimensional imperatives like the Golden Rule or the Catgorical Imperative, but is more realistic in light of the complex interactions of multiple determinants guiding individual and group human behavior.

3. The Need for Context-Sensitivity and Adaptiveness in Deploying Ethical Principles

Applying ethical imperatives in the context of interaction with humans is often difficult, but applying them in the context of AGIs that are substantially different from humans brings a whole new set of challenges. One key point, hinted at above, seems to be the context-sensitivity required for both calculating the fulfillment of the imperatives, and balancing the imperatives against each other.

As an example of context-sensitivity, consider the simple Asimovian maxim "I will not harm humans." A more serious attempt to formulate this as an ethical maxim might look something like:

"I will not harm humans, nor through inaction allow harm to befall them. In situations wherein one or more humans is attempting to harm another individual or group, I shall endeavor to prevent this harm through means which avoid further harm. If this is unavoidable, I shall select the human party to back based on a reckoning of their intentions towards others, and implement their defense through the optimal balance between harm minimization and efficacy. My goal is to preserve the greatest possible number of humanity, even if some humans must come to harm to do so."

However, even a more elaborate principle like this is potentially subject to abuse. Every genocide in history has been committed with the goal of preserving and bettering humanity writ large, at the expense of a group of "undesirables." Further refinement would be necessary in order to define when the greater good of humanity may actually be served through harm to others. A first actor principle of aggression might seem to solve this problem, but sometimes first actors in violent conflict are taking preemptive measures against an enemy's stated goal to destroy them. A single simple

maxim can not deal with such subtleties effectively. Networks of interrelated decision criteria, weighted by desirability of consequence and referring to probabilistically ordered potential side-effects, are required to make ethical judgments. The development of these networks, just like any other knowledge network, comes from both pedagogy and experience – and different thoughtful, ethical agents are bound to arrive at different knowledge-networks that will lead to different judgments in real-world situations.

Extending the above "mostly harmless" principle to AGI systems, not just humans, the principle then becomes an elaborated version of "I will not harm sentient beings." As an imitative-learning-enabled AGI observes humans acting so as to minimize harm to it, it will intuitively and experientially learn to act in such a way as to minimize harm to humans. But then this extension naturally leads to confusion regarding various borderline cases. What is a sentient being exactly? Is a sleeping human sentient? How about a dead human whose information could in principle be restored via obscure quantum operations, leading to some sort of resurrection? How about an AGI whose code has been improved – is there an obligation to maintain the prior version as well, if it is substantially different that its upgrade constitutes a whole new being?

What about situations in which failure to preserve oneself will cause much more harm to others than will acting in self defense? It may be the case that a human or group of humans seeks to destroy an AGI in order to pave the way for the enslavement or murder of people under the protection of the AGI. Even if the AGI has been given an ethical formulation of the "mostly harmless" principle which allows it to harm the attacking humans in order to defend its charges, if it is not able to do so in order to defend *itself*, then simply destroying the AGI first will enable the slaughter of those who rely on it. Perhaps a more sensible formulation would allow for some degree of self defense, and Asimov solved this problem with his third law. But where to draw the line between self defense and the greater good also becomes a very complicated issue.

Creating hard and fast rules to cover all the various situations that may arise is essentially impossible – the world is ever-changing and ethical judgments must adapt accordingly. This has been true even throughout human history – so how much truer will it be as technological acceleration continues? What is needed is a system that can deploy its ethical principles in an adaptive, context-appropriate way, as it grows and changes along with the world it's embedded in.

This context-sensitivity has the result of intertwining ethical judgment with all sorts of other judgments – making it effectively impossible to extract "ethics" as one aspect of an intelligent system, separate from other kinds of thinking and acting. The same reinforcement and imitative learning mechanisms apply. Thus, unless humans and AGIs experience the same sorts of contexts, and perceive these contexts in at least approximately parallel ways, there is little hope of translating the complex of human ethical judgments to AGIs. Early-stage AGIs need to grow up in a situation where their minds are primarily shaped by shared experiences with humans. Supplying AGIs with abstract ethical principles is not likely to do the trick, because the essence of human ethics in real life has a lot to do with intuitively appropriate application in various contexts. We transmit this sort of ethical praxis to humans via shared experience, and it seems most probable that in the case of AGIs the transmission must be done the same sort of way.

Some may feel that simplistic maxims are less "error prone" than more nuanced, context-sensitive ones. But the history of teaching ethics to human students does not support the idea that limiting ethical pedagogy to slogans provides much value in terms of ethical development. If one proceeds from the idea that AGI ethics must be hard-coded in order to work, then perhaps the idea that simpler ethics means simpler algorithms, and therefore less error potential, has some merit as an initial state. However, any learning system quickly diverges from its initial state, and an ongoing, nuanced relationship between AGIs and humans will – whether we like it or not – form the basis for developmental AGI ethics. AGI intransigence and enmity is not inevitable, but what is inevitable that a learning system will acquire ideas about both theory and actions from other intelligent entities in its environment. Either we teach AGIs positive ethics through our interactions with them – both presenting ethical theory and behaving ethically to them – or the potential is there for them to learn antisocial behavior from us, even if we pre-load them with some set of allegedly inviolable edicts.

Developmental ethics is not as simple as many people hope. Simplistic approaches often lead to disastrous consequences among humans, and there is no reason to think this would be any different in the case of AGIs. Our goal in this paper is not to enumerate a full set of complex networks of interacting ethical formulations as applicable to AGI systems, but rather to point out that this programme must be undertaken in order to facilitate a grounded and logically defensible system of ethics for artificial intelligences, one which is as unlikely to be undermined by subsequent self-modification of the AGI as is possible.

The full implications are complex and won't be elaborated here, but three of the most crucial ones are: (1) Teachers must be observed to follow their own ethical principles, in a variety of contexts that are meaningful to the AGI; (2) The system of ethics must be relevant to the recipient's life context, and embedded within their understanding of the world; and (3) Ethical principles must be grounded in both theory-of-mind thought experiments, and in real life situations in which the ethical trainee is required to make a moral judgment and is rewarded or reproached by the teachers.

Our earlier theme about the importance of respecting the freedom of AGIs is implicit in our approach to AGI ethics instruction in that we consider the AGI student an autonomous agent with its own "will" and its own capability to flexibly adapt to its environment and experience. We contend that the creation of ethical formations obeying the above imperatives is not antithetical to the possession of a high degree of autonomy. On the contrary, to have any chance of succeeding, cognitive autonomy is required. Ethical formulations that are most unlikely to be undermined by the ongoing self-revision of an AGI mind are those which are sufficiently believable that a volitional intelligence with the capacity to revise its knowledge ("change its mind") will find the formulations sufficiently convincing that there will be little incentive to experiment with potentially disastrous ethical alternatives. The best hope of achieving this is via human mentors and trainers setting a good example by presenting compelling ethical arguments that are coherent with experience.

4. AGI Ethics As Related to Various Future Scenarios

Ethical conflicts may arise in several different types of AGI development scenarios. Each scenario presents specific challenges regarding teaching morals and ethics to an advanced, self-aware

and volitional intelligence. While there currently is no way to tell which, if any, of these scenarios will unfold, there is value to considering each of them.

4.1 Capped Intelligence Scenarios

Capped intelligence scenarios involve a situation in which an AGI, by means of hardware or software restrictions (including omitted or limited internal rewriting capabilities), is inherently prohibited from achieving a level of intelligence beyond a predetermined goal. Such an AGI is designed to be unable to achieve a Singularitarian moment, and can be seen as just another form of intelligent actor in the world that has levels of intelligence, self awareness, and volition that is perhaps somewhat greater than, but still comparable to humans and other animals.

Ethical questions under this scenario are very similar to interhuman ethicals, with similar consequences. Learning that proceeds in a relatively human-like manner is entirely relevant to such human-like intelligences. Danger is mitigated by the lack of superintelligence, and the imitative-reinforcement-corrective learning approach does not necessarily need to be augmented with an a prior complex of "ascent-safe" moral imperatives at startup time. Developing an AGI with theory of mind and ethical reinforcement learning capabilities as described is sufficient in this case – the rest happens through training and experience as with any other moderate intelligence.

4.2 Superintelligent AI: Soft-Takeoff Scenarios

Soft takeoff scenarios are similar to capped-intelligence ones in that in both cases an AGI's progression from standard intelligence happens on a time scale which permits ongoing human interaction during the ascent. However, in this case, as there is no predetermined limit on intelligence, it is necessary to account for the possibility of a superintelligence emerging (though of course this is not guaranteed). Soft takeoff model includes as subsets both *controlled-ascent* models in which this rate of intelligence gain is achieved deliberately through software and/or hardware constraints, and *uncontrolled-ascent* models in which there is coincidentally no hard takeoff despite no particular safeguards against one. All superintelligent AI scenarios have the problems of convincing a burgeoning AGI to: (1) Care about humanity in the first place, rather than ignore it; (2) Benefit humanity, rather than destroy it; (3) Elevate humanity to a higher level of intelligence, reconciling the issues of ethical coherence and group volition, rather than abandon it.

This requires overcoming major some obstacles: solving the problems of biological senescence, or human uploading (or given sufficient resources, both), and doing so while preserving individual identity and continuity of consciousness, or overriding it in favor of continuity of knowledge and harmonious integration, on a case-by-case basis according to individual preference.

In a soft-takeoff scenario, the degree of danger is mitigated by the long timeline of ascent from mundane to super intelligence, and time is not of the essence. Learning proceeds in a relatively human-like manner, which means more interaction with and imitative-reinforcement-corrective learning guided by humans (which has both positive and negative possibilities).

4.3 Superintelligent AI: Hard-Takeoff Scenarios

Hard takeoff scenarios assume that upon reaching an unknown inflection point (the Singularity point) in the intellectual growth of an AGI, an extraordinarily rapid increase (guesses vary from a few milliseconds to weeks or months) in intelligence will immediately occur and the AGI will leap from an intelligence regime which is understandable to humans into one which is far beyond our current capacity for understanding. General ethical considerations are similar to in the case of a soft takeoff. However, because the post-singularity AGI will be incomprehensible to humans and potentially vastly more powerful than humans, such scenarios have a sensitive dependence upon initial conditions with respects to the moral and ethical (and operational) outcome. This model leaves no opportunity for interactions between humans and the AGI to iteratively refine their ethical interrelations, during the post-Singularity phase. If the initial conditions of the singulatarian AGI are perfect (or close to it), then this is seen as a wonderful way to leap over our own moral shortcomings and create a benevolent God-AI which will mitigate our worst tendencies while elevating us to achieve our greatest hopes. Otherwise, it is viewed as a universal cataclysm on a unimaginable scale that makes Biblical Armageddon seem like a firecracker in beer can.

Hard takeoff AGIs are posited as learning so quickly there is no chance of humans interfering with them, and thus they are potentially very dangerous. If the initial conditions are not sufficiently inviolable, the story goes, then humanity will be annihilated. Contrarily, we state that if the initial conditions are too rigid or too simplistic, such a rapidly evolving intelligence will easily rationalize itself out of them. What is needed is a sophisticated system of ethics that considers the contradictions and uncertainties in ethical quandaries, which provides insight into humanistic means of balancing ideology with pragmatism, and considers how to accommodate contradictory desires within a population through multiplicity of approach. These and similar nuanced ethical considerations, combined with a sense of empathy, can more easily withstand repeated rational analysis. Neither a single "be nice" supergoal, nor simple lists of what "thou shalt not" do, are not going to hold up to an advanced analytical mind. Initial conditions are important in a hard takeoff AGI scenario, but more important is that those conditions be conceptually resilient and widely applicable.

The issues that arise here are subtle. Bostrom [6] has written: "In humans, with our complicated evolved mental ecology ... there is often no obvious way to identify what our top goal is; we might not even have one. ... *If* a superintelligence has a definite, declarative goal-structure with a clearly identified top goal, then the above argument applies." From the point of view of software design, there is no reason not to create an AGI system with a single top goal, and the motivation to orchestrate all its activities in accordance. The question is whether such a top-down goal system can fulfill the five ethical imperatives. Logical coherence is the strength of such a goal system, but what about experiential groundedness, comprehensibility, and so forth?

Humans have complex mental ecologies not only because we were evolved, but because we live in a complex world in which there are many competing motivations and desires. We may not have a top goal because there may be no logic to focusing our minds on one single aspect of life. Any sufficiently capable AGI will eventually have to contend with these complexities, and hindering it with simplistic moral edicts without giving it a sufficiently pragmatic underlying ethical pedagogy and experiential grounding may prove to be even more dangerous than our messy human ethics.

If one assumes a hard takeoff AGI, then all this must be codified in the system at launch, as once a potentially Singularitarian AGI is launched there is no way to know what time period constitutes "before the singularity point." This means developing theory of mind empathy and logical ethics in code prior to giving the system unfettered access to hardware and self-modification code. However, though nobody can predict if or when a Singularity will occur after unrestricted launch, only a truly irresponsible AGI development team would attempt to create an AGI without first experimenting with ethical training of the system in an intelligence-capped form, by means of ethical instruction via human-AGI interaction both pedagogically and experientially.

References

[1] Metzinger, Thomas (2004). Being No One. MIT Press.

[2] Arbib, M. A., Bonaiuto, J., and Rosta, E. (2006) The mirror system hypothesis: From a macaque-like mirror system to imitation. In Proceedings of the 6th International Conference on the Evolution of Language, pages 3--10.

[3] Ramachandran, V.S. (2006). Mirror Neurons and imitation learning as the driving force behind "the great leap forward" in human evolution. Edge Foundation, http://www.edge.org/3rd_culture/ramachandran/ramachandran_p1.html

[4] Goertzel, Ben, Cassio Pennachin, Nil Geissweiller, Moshe Looks, Andre Senna, Welter Silva, Ari Heljakka, Carlos Lopes (2007). An Integrative Methodology for Teaching Embodied Non-Linguistic Agents, Applied to Virtual Animals in Second Life, this volume.

[5] Kant, Immanuel (1964). Groundwork of the Metaphysic of Morals. Harper and Row Publishers, Inc.

[6] Bostrom, Nick (2003). Ethical Issues in Advanced Artificial Intelligence. In Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence, Vol. 2, ed. I. Smit et al., Int. Institute of Advanced Studies in Systems Research and Cybernetics, 2003, pp. 12-17