Apparent Limitations on the "AI Friendliness" and Related Concepts Imposed By the Complexity of the World

Ben Goertzel
September, 2006

This document consists of a set of rough notes, analogous in nature to a blog entry. It is not a carefully-wrought formal essay. It may be replaced by one eventually. It is being posted now to encourage discussion.

Introduction

Suppose that, through advances in science and technology, we humans manage to create artificial intelligence systems that dramatically exceed ourselves in intelligence and pragmatic power. An important question then arises: How will these AI's act toward us humans? Eliezer Yudkowsky, in his various online writings (see links at www.singinst.org), has introduced the term "Friendly AI" to refer to powerful AI's that are beneficent rather than malevolent or indifferent to humans.[1] On the other hand, in my prior writings (see the book *The Path to Posthumanity* that I coauthored with Stephan Vladimir Bugaj; and my earlier online essay "Encouraging a Positive Transcension"), I have suggested an alternate approach in which much more abstract properties like "compassion", "growth" and "choice" are used as objectives to guide the long-term evolution and behavior of AI systems. This essay further explores this area of thinking, giving some arguments as to why seeking to create advanced AI systems embodying AI Friendliness and other highly specific goals may be an un-fruitful pursuit.

My general feeling, related here in the context of some specific arguments, is not that Friendly AI is a bad thing to pursue in any moral sense, but rather that it is very likely to be unachievable for basic conceptual reasons. I don't claim to have proved this incontrovertibly, just to have given some suggestive and (to me) intuitively convincing arguments. The arguments given here support my prior contention that it is more feasible and potentially fruitful to think about instilling AI's with general properties than highly specific properties like Friendliness (to humans). This point is quite independent of the issue of whether Friendliness (to humans) is a "better" or "worse" goal for an AI to preserve through its evolution than more general goals like compassion, choice and growth. My point here is not about what is the best kind of goal, it's about what kind of goal is more likely to be achievable.

My point is not just that "creating Friendly AI is hard." There is a difference between things (like creating powerful general AI in the first place) that are hard but achievable with effort and sufficiently advanced technology, and things that are

[1] Our usage of the term may not agree precisely with Yudkowsky's current usage (and his usage has shifted somewhat over the years), but we believe the spirit is the same.

fundamentally so difficult they may well never be achieved. I suggest that Friendly AI may fall into the latter category. Of course, we can't really say with any certainty that anything will be unachievable post-Singularity. But if we're going to try to reason about the post-Singularity future, we may as well listen to what our logic tells us. And what logic and experience tell us is that some things, like faster-than-light travel and highly advanced, guaranteeably Friendly AI, may well wind up being impossible even after the Singularity. And if they do become possible in the future, it may be in the context of other ideas, constraints and phenomena that are completely obscure to us at this time, thus impossible to presently reason about.

My goal in this essay is to explore some particular aspects of the difficulty of creating Friendly AI, which ensue not from the subtleties of AI design but rather from the complexity of the notion of Friendliness itself, and the complexity of the world in which both humans and AI's are embedded. The arguments I present here are conceptually fairly simple ones; my goal here is to be as clear as possible about some relatively simple points rather than to make highly subtle arguments. Subtle and powerful arguments about Friendly AI may someday be possible, but if so they will be built on the foundation of a clear understanding of the simpler aspects.

As already noted, the overall gist of my comments here is that Friendliness may not be a very useful notion for thinking about future AI's and their benevolence or otherwise. The concept is slippery in disturbing ways, and poses requirements that are most likely unachievable given the nature of reality. Furthermore, this conclusion is not closely tied to any particular interpretation of what "Friendly" means – rather, it has to do with problems associated with trying to pre-specify things so that advanced AI's will be compelled to do *any* very specific sort of thing. The conclusion of the arguments presented here is that compelling, in advance, advanced AI's to do any very specific sort of things is probably not a plausible enterprise. This may sound pessimistic, but it's not entirely so. I do think it may be possible to create advanced AI's in such a way as to encourage them to possess certain highly general properties – but not properties as specific as "benevolence to humans" or other sorts of "AI Friendliness" type qualities. Rather, it may be possible to create them in such a way as to encourage them to display general properties like compassion, growth and choice. I don't know for sure that this is the case, but the arguments given here don't rule such a thing out, in the same way that they apparently rule out AI Friendliness as a plausible thing.

As a preliminary to presenting my arguments, I will first distinguish two aspects of Friendliness, which I call action-based and outcome-based. An action-based Friendliness criterion is one that rates the actions of an agent (for instance, an AI or a human), in a particular context, and judges whether they appear "Friendly" or not. An outcome-based Friendliness criterion is one that rates the actions of an agent, in a particular context, and judges whether they are going to lead to a "Friendly" outcome or not, over some future period of time (perhaps a fixed future interval of time). Put crudely, an action-based criterion judges whether an agent appears to be "trying to do the right thing," whereas an outcome-based criterion judges whether an agent is actually doing the right thing in terms of the consequences of its actions. In the mind of a reasonably consistent and rational agent, action-based criteria and outcome-based criteria are connected via beliefs: the agent will belive that if actions fulfilling its action-based criteria are carried out, this renders it differentially likely that outcomes matching its

outcome-based criteria are fulfilled (for instance: significantly more likely than if those actions were not carried out and instead some random actions were carried out).

In these terms, the basic arguments I present here regarding Friendliness are as follows:

- Creating accurate formalizations of current human notions of action-based Friendliness, while perhaps possible in the future with very significant effort, is unlikely to lead to notions of action-based Friendliness that will be robust with respect to future developments in the world and in humanity itself
- The world appears to be sufficiently complex that it is essentially impossible for seriously resource-bounded systems like humans to guarantee that *any* system's actions are going to have beneficent outcomes. I.e., guaranteeing (or coming anywhere near to guaranteeing) outcome-based Friendliness is effectively impossible. And this conclusion holds for basically any highly specific property, not just for Friendliness as conventionally defined. (What is meant by a "highly specific property" will be defined below.)

The arguments I present here, along these lines, are not rigorous in the manner of mathematical proofs, but attempt to be carefully-reasoned in the manner of classical philosophy arguments.

(Difficulties With) the Formalization of Friendliness

First of all, one comment-worthy aspect of the notion of AI Friendliness is its incredible slipperiness. This is a side point to the main arguments being presented here, but it needs to be dealt with. Any formal definition of Friendliness one poses turns out to have glaring exceptions. An early and lucid illustration of this problem was given in Jack Williamson's classic novel *The Humanoids*, in which mildly superhuman AI robots are created and supplied with the goal system "To Serve and Protect, and Guard Men from Harm." Lo and behold, the robots interpret this to encompass guarding men from psychological harm, such as thinking troubling thoughts or working on overly difficult research; and they proceed to compel all humans to have happy minds, via forcible administration of appropriate drugs if necessary. In the same vein, any definition one poses seems to be susceptible to exceptions. As Williamson demonstrated, keeping humans happy is not a good Friendliness criterion, as this would be too easily achieved via administration of happy-drugs. Alternately, any Friendliness involving respecting human freedom is mighty tricky to make precise, since none of us fully understands what "freedom" is, politically, neurally or morally. Posing Friendliness criteria and poking holes in their vagueness or incompleteness is an amusing pastime, but rapidly becomes too easy. There are of course individual variations regarding what is considered Friendly, but this turns out not to be the biggest issue: a deeper problem is that formalizing any individual notion of Friendliness is incredibly difficult. And, this problem seems to occur for both action-based and outcome-based Friendliness.

However, the slipperiness of Friendliness does not strongly distinguish it from other commonsense human concepts. The Friendliness-formalization issue is somewhat similar to the problems that early logic-oriented AI theorists found in trying to

encapsulate everyday notions like "cup" or "over" in terms of logic formulas. A few logic formulas may suffice to capture 80% of what makes an object a cup, but then after that, each logic formula one adds to the definition captures a rapidly decreasing portion of the everyday notion of cup-ness – and in the end, the conclusion is that abstract logical formulations are not the right way to capture everyday concepts. The problem is not with the logic aspect but with the abstractness aspect: real human concepts like "cup" and "over" are complex mixes of abstract formulas with specific exemplars and other sorts of patterns. Cup-ness may be expressible in terms of logic, but only using massively complex logic formulas that embody references to various exemplar cups and various exemplars of othe related objects. Similarly, encapsulating any reasonable commonsense notion of Friendliness in a set of compact logical formulas seems not to be possible: not because Friendliness is intrinsically unformalizable, but because the human notions of beneficence and morality are massively complex combinations of abstractions, comparisons to exemplars, analogies and other mental patterns.

So: defining Friendliness formally is extremely difficult. However, this doesn't mean it is necessarily impossible. It seems plausible that a mildly superhuman AI – or a specialized AI system with infrahuman general intelligence -- could be tasked with creating a massive logic formula capturing the action-based and outcome-based notions of Friendliness implicit in some particular human's world-view. Just as it will eventually be possible to create an AI capable of imitating a particular human's judgments of what is or is not a cup, similarly it will eventually be possible to create an AI capable of imitating a particular human's judgments of which actions and outcomes are or are not Friendly.[2]

So, let's suppose that this problem were solved, and we had a formalization of a human being's action-based and outcome-based notions of Friendliness. The next question is: How far would this take us? How useful would this be? My argument is that it wouldn't really be very useful. The difficulty of formalizing Friendliness is severe but probably not insurmountable – but it doesn't matter much, because even a fully formalized version of some human or human group's version of Friendliness wouldn't be very valuable for the wise Singularitarian AI designer.

The Near-Uselessness of Action-Based Friendliness Criteria

In the following section – the meatiest one of the essay -- I will address the question of outcome-based notions of Friendliness, arguing that given the complexity of the world, it's probably impossible to guarantee that any nontrivial outcome-based Friendliness criterion will be fulfilled, no matter how well formalized it is. But what about action-based Friendliness criteria? What about specifying a goal that may be

---

[2] Note that this is superficially related to Yudkowsky's notion of Coherent Extrapolated Volition, in which a super-powerful specialized AI system is posited as a route to figuring out "what humans would want if they were the people they wanted to be." However, the parallel does not run very deep. Both suggestions involve creating an AI to help formulate the goal of Friendliness, but, CEV is a much more ambitious and trickier proposal.

fulfilled by the fact of an AI acting a certain way (based on what it perceives in its environment), regardless of the actual empirical consequences of the AI's actions?

It might be possible to architect an AI so that, even in its future incarnations as they interact with future environments, the AI will always carry out actions fulfilling some pre-specified criterion (some formalization of action-based Friendliness). Whether this is possible or not is unclear, but impossibility is not obvious. What does seem clear, however, is that this kind of ongoing action-criterion-based Friendliness is almost useless except in the case of a nearly-constant environment and a non-evolving AI system.

The problem is that what really matters are outcomes, not actions! The connection between outcomes and actions is beliefs; and for any rational agent, as new information about the world is received, or new cognitive abilities are achieved, beliefs change. If the world changes significantly, new information about the world is going to come into the mind of any intelligent system perceiving the world, and change its beliefs about which actions are most likely to bring about appropriate outcomes. Furthermore, even if most of the world remains basically the same, if an AI system increases significantly in intelligence, it is likely to understand new connections between actions and outcomes that were previously opaque to it – and, once again, change its beliefs about which actions are most likely to bring about appropriate outcomes.

A superhuman AI system compelled to confront the uncertain future by carrying out an action plan determined to fulfill a non-outcome-based criterion designed by human beings (who have relatively low intelligence and relatively little experience of the world), will be a mind that is doomed to carry out an action plan that it knows is nowhere near maximally likely to achieve the ultimate outcomes desired by itself or its creators.

And this leads us right back to outcome-based Friendliness. Rather than specifying exactly what criteria an AI's actions must fulfill, why can't we specify what outcomes an AI's actions are supposed to lead to? What I'll argue in the next section is that this kind of approach can't work either, unless the Friendliness criterion (used to assess outcomes) is extremely simplistic. We argue that, for any reasonable Friendliness criterion, the complexity of the world renders it essentially impossible to guarantee that any system (superhuman or not) will obey it.

A Semi-Formalization of the Problem of Guaranteeing AI Friendliness

In this section I present a semi-formalized treatment of the problem of guaranteeing outcome-based AI Friendliness. The use of some formal notation and language is mainly for sake of clarity, as natural language is intrinsically ambiguous and often leads to unnecessary confusion when topics as multidimensionally tricky as Friendly AI are concerned. The notation and language should not lead the reader to believe that I have proved any deep mathematical theorems about AI Friendliness: I have not, at this stage, although such a project is indeed of interest to me.

In fact, most of the discussion in this section is not explicitly about Friendliness, but is about any predicate F specifying certain properties. The idea is that, no matter how one defines Friendliness, if one's definition satisfies certain properties then the arguments given here are applicable.

So: suppose we have some AI system S which is part of some world W; and assume that the state of the world W at any time has a finite description. Suppose we

have some ternary property F, defined as a ternary predicate F(W,I) that can be evaluated in terms the set of world-states W(t) corresponding to time points t within some interval I. By a ternary predicate is meant a predicate whose value on any argument is either True, False or Neutral. F is intended to evaluate some property of S as manifested via S's interaction with the rest of the world, but it can be defined as a predicate with argument W because S is assumed part of W. Finally, suppose we have a monitor system M, whose job is to study S and W over time, and determine whether F is fulfilled or not.

Now, suppose that W has high predictive complexity, in the following sense. Suppose that in order to predict whether or not W(t+s) will fall within a sphere of radius r or less, based on any (even the most useful) b bits of knowledge of W(t), requires computing power C(t,s,r,b). Suppose that C(s,r,b) is monotonically increasing in s, and monotonically decreasing in r and b. (Furthermore, in practice we may suppose that the dependence of C on s is very rapidly increasing, and the dependence on r and b is very rapidly decreasing.)

Then, suppose there are two world-states x and y so that d(x,y)>=r, and so that if X is a world-state series over (t,s) for which W(s)=x then F(X)=True, whereas if X is a world-state series over (t,s) for which W(s)=y then F(X)=False. Then, the only way M can guarantee (prove) that W will fulfill F during the time period (s,t) is if M has computing power greater than or equal to C(t,s,r,b).

Suppose b is equal to the number of bits that M can access from its memory within the time-span s-t. Then we may restate the bound on M's computing power as C(t,s,r).

Now, suppose that the world is *dynamically complex*, in the sense that as s increases and/or r decreases, the amount of computing power required for M, operating around time t, to predict the applicability of F during (t,s), increases very rapidly.

Or to put it differently, if the world is dynamically complex in the sense intended here, then given a particular fixed F (and a fixed r associated with F), the amount of computing power required of M in order for M to predict the applicability of F during (t,s) increases incredibly rapidly with s.

I have not proved this, but I suspect that dynamical complexity in the sense described here may closely related to the mathematical notion of the "topological entropy" of a dynamical system.

So, if the world is dynamically complex and we have a particular function F that we are concerned about, then unless this F corresponds to an extremely (unrealistically) crude partition of the set of world-states, it is not going to be possible for a realistic monitor M to assess whether F will be fulfilled.

Setting aside the assumption of dynamical complexity, we may also draw some other conclusions. In general, what the above line of argument implies is that the only hope M has of guaranteeing F's truth during (t,s) is if the world-state-series F judges True are very different from the world-state-series F judges False, with a broad band of Neutral states inbetween them. On the other hand if F embodies subtle judgments between True and False state-series then r will necessarily be small and the computing requirements imposed on F will be unfulfillable.

Conceptually, the conclusion from the prior paragraph is that if the applicability of criterion F to a reasonably complex, world-interacting AI system S is going to be provably established by a finite monitor system M (like a human, or a simpler AI), then

a) F had better involve a very crisp distinction of True versus False world-state-series.

b) the world W had better not display significant dynamical complexity in the above sense

Point b is a tricky one in that the world of classical and general relativistic physics does appear to display dynamical complexity in the above sense[3], yet the world of quantum physics may not[4]. So it might be that a highly powerful monitor using quantum physics for prediction and a very crisp F could create a powerful AI system S and prove that it would fulfill F over time as it developed. But this is of course quite hypothetical as our understanding of physics is incomplete and it might happen that the (as yet unknown) correct unification of quantum theory and general relativity reveals that world does indeed possess significant dynamical complexity.

Summary and Conclusion

I have argued that, while difficult, the problem of formalizing a particular human's implicit internal notion of AI Friendliness may not be impossible – and may be achievable with the help of future AI systems. Further, I have distinguished two different types of possible AI Friendliness criteria: action-based and outcome-based. I have argued that action-based Friendliness criteria are not desirable because they do not allow AI's to use the benefits of their increased intelligence and experience to figure out how better to achieve desired outcomes using modified actions plans. And I have argued that outcome-based Friendliness criteria are almost surely not tractably specifiable, given the complexity of the world. None of these arguments are mathematically formal, hence none should be considered ironclad and irrefutable. Yet, I do think these arguments are conceptually evocative and meaningful, and I think what they suggest is that in a strong sense, "AI Friendliness" is probably not a particularly useful way to look at the problem of encouraging beneficial outcomes in the context of posthuman AI's. Alternate perspectives are going to be required as we confront the reality of AGI's and the Singularity.

And, what kinds of alternate perspectives? The arguments of the prior section suggest one possible way out of the issues that Friendly AI has with the complexity of the world. This is to look at predicates F that are so simple they very clearly distinguish True from False: predicates involving distinctions that are more simple and basic to the universe than any human-based notion of Friendliness could possibly be. This suggests

---

[3] I have not proved this formally but it seems very likely given the mathematical results of "chaos theory"

[4] This relates to difficulties with the notion of "quantum chaos": while quantum systems may appear chaotic in their "classical limits," the equations of quantum physics do not allow chaos in the sense of exponential sensitivity to initial conditions, and this may place limits on the extent to which quantum systems may be "dynamically complex" in the sense loosely described above.

an interesting area for investigation: whether general notions like compassion, growth and choice could possibly be shown to possess the property of being "easily distinguishable predicates" in the above sense. That is, it may be that the set of uncompassionate world-states (in some appropriate definition of "compassion") is so clearly distinguishable from the set of compassionate world-states, that it is possible circumvent the world-complexity-based limitations noted above. Of course, this is just a speculation, but I consider it an interesting direction for investigation.

What seems clear to me, however, is that the notion of AI Friendliness is deeply troubled, and needs to be replaced with something new that is more friendly to the actual, complex nature of the universe. This is a pragmatic rather than a moral point: whether or not you think human-Friendliness is a "better" goal than e.g. compassion, you still need to ask yourself whether it is a plausibly achievable goal.