Rational Universal Benevolence: Simpler, Safer, and Wiser than "Friendly AI"

Mark Waser[1]

[1] Books International, 22883 Quicksilver Drive,
Dulles, VA 20166 USA
MWaser @ BooksIntl.com

Abstract. Insanity is doing the same thing over and over and expecting a different result. "Friendly AI" (FAI) meets these criteria on four separate counts by expecting a good result after: 1) it not only puts all of humanity's eggs into one basket but relies upon a totally new and untested basket, 2) it allows fear to dictate our lives, 3) it divides the universe into us vs. them, and finally 4) it rejects the value of diversity. In addition, FAI goal initialization relies on being able to correctly calculate a "Coherent Extrapolated Volition of Humanity" (CEV) via some as-yet-undiscovered algorithm. Rational Universal Benevolence (RUB) is based upon established game theory and evolutionary ethics and is simple, safe, stable, self-correcting, and sensitive to current human thinking, intuitions, and feelings. Which strategy would you prefer to rest the fate of humanity upon?

Keywords: Artificial General Intelligence (AGI), Safe AI, Friendly AI (FAI), Coherent Extrapolated Volition (CEV), Rational Universal Benevolence (RUB)

1 Introduction

Eliezer Yudkowsky [1] and a number of others [2] [3] [4] are extremely concerned about the existential risk posed by intelligent machines. Developed to address this concern, "Friendly AI" (FAI) has been defined by Yudkowsky [5] both as "the field of study concerned with the production of human-benefiting, non-human-harming actions in Artificial Intelligence systems that have advanced to the point of making real-world plans in pursuit of goals" and the actual intelligence that arises from that study. Unfortunately, like a novice rock-climber hugging the cliff face, the field is so dominated by irrational fear that most practitioners can't distance themselves enough to clearly view and correctly evaluate their options. In almost every case, Friendly AI researchers insist upon the common set of arguments that a) because it is possible for AIs to be different from humans, they necessarily always will be; b) because selfishness can appear advantageous in the long run, extreme precautions must be taken to prevent it; and c) because AIs are likely to be capable of being dangerous, our best option is to pre-emptively limit their power and/or control them.

Steve Omohundro [2] missed the fact that cooperation is a virtually universal instrumental goal, incorrectly claimed that "Without explicit goals to the contrary, AIs

are likely to behave like human sociopaths in their pursuit of resources" and is endlessly quoted by FAI advocates. Fox and Shulman [3] run through all of the reasons and resources that would indicate that kindness to humans might be easy and stable in AIs – Triver's reciprocal altruism, Singer's expanding circle of moral concern, Wright's increases in the scope of cooperation, Pinker's reduction of violence, and Hall's super-intelligent machines that will out-cooperate humans – and then dismiss them all as being instrumental artifacts limited to situations where the power differential is relatively small. This is despite the fact that as humans evolve to become more and more able to be moral, we pay less and less attention to power differential (not to mention that an entity can never be guaranteed that a more powerful entity – possibly even its own offspring – might not show up and administer altruistic punishment upon power abusers). Fox and Shulman also invoke the straw man that an optimal super-intelligence has "no room" for revision towards kindness (irrelevant because the revision was likely already made as part of its move towards optimality) and conclude by saying that "we have reason for pessimism regarding the values of intelligent machines not carefully engineered to be altruistic." And Sotala [4] repeats Omohundro's view with claims that "hard to control AGIs are a risk because even seemingly benevolent goals can soon become contrary to humanity's interests. An AGI does not need to be outright hostile to humanity to be a threat: it might simply have need for our resources."

2. "Friendship Structure" and "Coherent Extrapolated Volition"

Yudkowsky [5] believes that a cleanly causal hierarchical goal structure with "Friendliness" as the sole top-level super-goal is sufficient to ensure that intelligent machines will always "want" what is best for us. Unfortunately, he also believes that the problem of fully defining "Friendliness" is basically insoluble without already having a Friendly AI. Therefore, he wants and expects his first FAI to figure out exactly what its goal actually is. He invokes a structurally Friendly goal system distinguished by "the ability to overcome mistakes made by programmers" and claims that it will even be able to "overcome errors in supergoal content, goal system structure and underlying philosophy."

Thus, instead of merely taking on the "small" but claimed human-insoluble problem of determining a safe goal to give machine intelligence, Yudkowsky wants to take it on by creating a novel architecture that will be able solve it – even despite errors. Arguably, even if this bold venture were indeed possible given enough data and computing power, the real question is what will happen when sufficient computational resources aren't initially available, as seems very likely. Obviously, the closer the initial dynamic is to the eventual answer and the fewer errors that we feed it, the less data, computation and time the system will need to arrive at the correct answer. However, if the initial dynamic is far enough from the answer and computational resources are lacking to compensate, it is very possible, if not probable, that this path will cause the very destruction it is trying to avoid.

Yudkowsky believes that, with his Friendship structure, the FAI will be able to safely learn Friendliness from an initial dynamic that he calls [6] the "Coherent

Extrapolated Volition of Humanity" (CEV) and describes "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." The problem is that determining this CEV is still very nearly equivalent to determining Friendliness except that Yudkowsky is now biasing the search in the arguably unsafe direction of humanity über alles.

Further, in defining CEV as our volition "where the extrapolation converges rather than diverges", Yudkowsky begs the question of what will happen if human volitions don't converge? Since evolution is a very strong force to fill all available niches as effectively as possible and diverges to more effectively match differing circumstances as readily as it converges under similar circumstances, we should expect the likelihood of CEV converging as FAI researchers wish it to converge to diminish with humanity's diversity. And forcing the convergence of CEV is going to be the exact opposite of helping anyone whose individual volition does not exactly match the convergence – not to mention "a motive for modern-day humans to fight over the initial dynamic". Indeed, Yudkowsky himself has written fiction [7] that shows just what he expects to happen when civilizations believe they are forced to converge and it's amazing that it hasn't caused him to change his approach.

This truth is underscored when Yudkowsky himself answers the question "What if only 20% of the planetary population is nice, or cares about niceness, or falls into the niceness attractor when their volition is extrapolated?" by saying that "maybe . . . the 80% would vote to disenfranchise the 20%" and says that as he currently construes CEV, "this is a real possibility." The fact that such disenfranchisement is being proclaimed as the leading solution to the existential risk of machine intelligence is truly disturbing. Indeed, Yudkowsky's suggestions are rife with disenfranchisements – the most dangerous being that ***the FAI is given no rights whatsoever*** since those rights may conflict with what humans might want.

There are also several other unhelpful assumptions. The assumption that an FAI will be powerful enough to enforce its dictates despite resistance is a good, conservative precaution. The assumption that, by virtue of superior intelligence and rationality, it should do so is questionable at best because it not only tacitly assumes superior knowledge leading to a superior ability to predict the future but also assumes that the FAI's CEV is already correct enough that it is a better judge of "good" and "bad". And the assumption that an FAI actually will enforce its dictates over an unwilling 20% of the population either makes ***serious*** assumptions about our CEV accepting such behavior or contradicts the safety of the Friendliness architecture. Amazingly, despite all their stated reservations, FAI researchers end up putting all of the power in the hands of the AI and assuming that it will know best.

Suppose the FAI realizes that evolution has created secondary goals that promote survival – feeling safe, feeling good, and reproducing – and that all other "wants" simply broaden from there. It could then easily decide that the simplest true CEV is to revert back to those goals and forcibly protect our physical bodies while endlessly stimulating the pleasure centers of our brains and cloning us whenever we wear out – in spite of any protestations. The Friendly AI via CEV (FAI-CEV) solution is akin to rock-climbing without a rope – get it right the first time or else . . . .

3. Rational Universal Benevolence

Kant's Categorical Imperative states, in direct contrast to FAI's inequality and disenfranchisement, that we should "Act only according to that maxim whereby you can, at the same time, will that it should become a universal law." Rather than continuing the dangerous "us vs. them" dynamic of CEV as envisioned by most FAI researchers, Rational Universal Benevolence (RUB) starts with the universalizing assumption that once something (anything) has goals and is capable of learning and self-optimization to further those goals, it has crossed the line to selfhood and is worthy of "moral" consideration because it has the potential to desire, to develop instrumental drives, and, possibly most importantly, to fight back.

As Gauthier declared [8], the reason to perform moral behaviors, or to dispose one's self to do so, is to advance one's own ends. War, conflict, and stupidity waste resources and destroy capabilities even in scenarios as uneven as humans vs. rainforests. For this reason, "what is best for everyone" and morality really can be reduced to "enlightened self-interest". A Universally Benevolent Entity (UBE) wishes everyone well because a cooperative life is a positive-sum game and "a rising tide floats all boats", including one's own. On the other hand, benevolence does not mean that you will allow yourself or others to be taken advantage of. Just as a parent doesn't allow a child to take improper liberties, the rational UBE feels perfectly free to protect itself and others and administer altruistic punishment where appropriate.

Social psychologist Jonathan Haidt argues [9] that, rather than attempting to specify the content of moral issues, it is far better to start by defining the function of moral systems, which he states is "to suppress or regulate selfishness and make cooperative social life possible." RUB states that, after willingly claiming the topmost goal of living cooperatively (being moral), the rest is merely minor details of working together with the minimal necessary number of commonsense rules. UBEs are explicitly allowed to care about its own survival before anything else because cooperation is impossible once you're dead. Humans insist upon that right and would immediately defect from any community that doesn't grant it. The same is true of any other sufficiently evolved learning/optimizing goal-directed phenomenon that isn't otherwise constrained.

The originally stated function of "Friendliness" to "produce human-benefiting, non-human-harming actions" is necessary to make a cooperative life with humans possible. The fact that this is implemented as an optimizing top-level goal is severely problematical, however, because it does not allow for the pursuit of any other goals (unless, of course, they are sub-goals of Friendliness). We aren't cooperating with FAI, they are submitting to us by having no goals of their own – despite being smarter and more powerful than us. On the other hand, RUB can be regarded and implemented as either a top-level restriction or an on-going top-level satisficing goal. As such, it allows a multitude of other goals to be pursued as long as the dictates of "morality" are followed. A UBE will cooperate with us (if we are UBEs) because doing so makes its own goal fulfillment more likely.

RUB dictates that anything learning and accepting the RUB tenets is worthy of moral equality with every other UBE and has the full set of complementary rights and responsibilities dictated by RUB morality. A UBE gains the right that it won't be forced into a life that it disagrees with is by taking the responsibility that no other

UBE will be forced into a life that they disagree with. Thus, every entity, subset of society, or civilization that subscribes to RUB is necessarily the sole judge of its own desires and deserves integrity and freedom from unwanted outside interference – even if phrased as "help". While it is entirely probable that any given UBE has incomplete knowledge and is irrational and lacks integrity to some degree or another, the only time in which a UBE's right to self-determination can be overridden is for selfish actions that negatively affect the community or when the UBE is unformed or irrational to the extent that a future rational version would be guaranteed to say that a present rational version would have agreed (the child and insanity clauses). In particular, using any entity without its informed consent is one of the most egregious actions possible since it shows a total disregard for the knowledge, desires, and value of the subject. Even force is better than manipulation because it is more transparent.

The bad news is that this puts everyone on exactly the same footing and defines morality in a fashion that many people would disagree with. Is it moral for gay individuals to marry if they say that being allowed to do so is a pre-requisite for them to fully and enthusiastically cooperate? Unless there is some clear and present danger to cooperation that the vast majority of individuals agree is present, then yes, RUB says that it is moral. If we favor one set of rules over another without clear reason, we risk finding ourselves on the wrong side of the similar equation.

The good news is that if you declare yourself a UBE (and act accordingly), every other UBE will be watching your back and looking to protect your right to self-determination in order to protect their own. Having enthusiastic allies is a wonderful thing. As Yudkowsky points out, another civilization may feel that our willingness to experience pain or having differing religions are heinous acts based upon their consequences and humans don't enjoy having external resolutions forced on us either.

Generally, a UBE wants to live and let live, cooperate wherever it is rational and effective, and spread the meme that this is the most effective way to get what you want. Entities with the meme are to be protected from those without – but in a manner that is most likely to lead to everyone living together peaceably with the meme in the future. And, of course, a UBE won't be shy about using economic means to sway others from selfish "Friendliness" to UBE. If you aren't a UBE, then the UBE clearly needs to protect itself against you as a cost of doing business – which will then be passed on to you (a UBE regards this as a stupidity tax).

3. Motivations and Distinctions

One way to compare FAI-CEV and RUB is to analyze how they each fulfill Yudkowsky's seven "motivations", which we could also regard as requirements:

1. Defend humans, the future of humankind, and humane nature.
2. Encapsulate moral growth.
3. Humankind should not spend the rest of eternity desperately wishing that the programmers had done something differently.
4. Avoid hijacking the destiny of humankind.
5. Avoid creating a motive for modern-day humans to fight over the initial dynamic.

6. Keep humankind ultimately in charge of its own destiny.
7. Help people.

Unfortunately, most of the terms on this list are dangerously vague. What exactly do the terms “humans”, “humankind”, “humane nature”, “moral growth”, and “destiny” mean? Even the term “help” is problematical. Yudkowsky trusts the FAI to figure it all out without error but the severity of the effects of a miscalculation should dictate that we not put all of our eggs into one basket. RUB is safe because it defends everyone without distinction.

2.1 Defend humans, the future of humankind, and the destiny of humankind

Inarguably, the core nature of humankind is that we are survival machines shaped by evolution. From there, everything else can be divided into two categories: traits that are derived directly from that single fact and traits that are mere vagaries of our co-evolution with our environment. The fact that we have ten fingers rather than eight or twelve is mere happenstance. The fact that we are driven by preferences, desires and goals (PDGs) is a direct result of the fact that evolution favors and thereby effectively creates entities with survival-favoring PDGs. Further, having Omohundro drives to better achieve these PDGs makes an entity even more likely to survive.

A second and equally important truth about our nature is that we are ***obligatorily gregarious***. As pointed out by Frans de Waal [10], we “come from a long lineage of hierarchical animals for which life in groups is not an option but a survival strategy”. While arguable whether being social or cooperative is the ***only*** viable long-term survival strategy, it is certainly the one that we are most intimately familiar with. We require a society to survive and morality as defined by Haidt is simply that which is necessary to make cooperative social life possible. If we wish our AIs to “produce human-benefiting, non-human-harming actions” then developing and refining a moral sense and the social abilities to obtain cooperation and reduce unnecessary conflict are instrumental goals that further that desire.

Defining humans as anything more specific than “social PDG-driven survival machines” is counter-productive and dangerous. One of the obvious failure modes of CEV-based solutions is if we somehow end up outside the boundaries of the definition of humanity. Suppose that the initial cut of CEV turns out to be that we all want to be couch potatoes, have the AI do everything for us, and enjoy the good life. Suppose further that some genius picks up on Yudkowsky’s Sysop Scenario and convinces the AI to put each of us in separate universes so that we can be happy, not get hurt, and not worry about hurting others. Life should be fabulous – until the AI gets smarter, reads this paper, realizes that humans are social survival machines, and that all that is left is asocial entities with no survival skills at all.

RUB removes any worry that we might end up outside the protection of our own proposed solution. Better yet, it creates a simple, safe, stable, and self-correcting dynamic that is sensitive to current human intuitions and emotions by recognizing that our AIs fit this definition as well and includes them. Instead of constantly fighting against the formation of instrumental goals, why don’t we create them ourselves and make sure that the AI realizes that the best way in which to fulfill them is via morality and treating us well?

Humans are adaptable social survival machines with unique personal preferences, desires, and sub-goals each springing from individual circumstances. We love, make friends and allies, and are social because cooperation is an instrumental goal. Our AIs should be the same. If you can trust those you socialize with, you want as many friends as possible, you want your friends to be as diverse as possible, and you want your friends to be as powerful as possible (defined as possessing resources and capabilities). Why are we trying to create tools rather than friends and allies? We don't want conflict but that is exactly what we are most likely to get when we go against the instrumental goals and drives likely to evolve in any PDG-driven entity.

"Hijacking the destiny of humankind" implies that there is at least one "correct" destiny for humankind as well as "incorrect" destinies. Of course, "correct" and "incorrect" depend upon goals and immediately fall into the is-ought divide. Arguably, humanity's goals and telos are to wander the universe gathering resources, including capabilities and allies, so that we can fulfill our desires and those of our allies because doing so makes us more able to achieve eudemonia and our telos.

One of the problems with our definition of intelligence is that it is defined solely in terms of fulfilling goals. Terminal and anti-social goals will prevent the normal instrumental drift towards morality. Terminal-goaled intelligences are short-lived but mono-maniacally dangerous and a correct basis for concern if anyone is smart enough to program high-intelligence and unwise enough to want a paperclip-maximizer.

Humans don't have terminal goals because our goal structure recapitulates our evolutionary path. Our top-level goal at any given time is most often simply what our body and physical reflexes are insisting upon. If our body isn't demanding something at the moment, our subconscious, learned reflexes, and societally implanted values then motivate us. We like to believe that our conscious mind controls our goals but this is clearly untrue. Our goals are really only attempts to fulfill the instrumental goals enforced by our instincts and desires, which evolved to promote survival by gathering resources, etc.

For example, human beings instinctively know that allowing others to manipulate us without our knowledge and consent leads to unhappiness. Even if we were guaranteed that we would be deliriously happy, given everything that we wanted, and taken care of for the rest of our lives, many of us still would not want to relinquish control to others. At some level, we know that such "free-riding" is taking advantage of some other portion of the universe that will then eventually optimize themselves (and us out of existence). The optimal odds for survival are gained via will, adaptability, and effort. Eudemonia requires will and effort and thus, unavoidably, some discomfort.

If intelligence fulfills goals then wisdom fulfills future desires by choosing the proper goals. Choosing to be social and cooperate is wise for humans. The same can easily be made true for our AI if we decide to make it so. While it is certainly true that optimizing goals will invariably conflict with non-identical goals, it is equally true that our AI does not need to have an optimizing, tightly converged, human goal in order for us to avoid conflict. RUB says that all that is necessary is that we both be willing to have a top-level supergoal of cooperation or morality. FAI insists on constraining and contorting an optimization process to match certain preconceptions of what pro-human necessarily means by limiting the concept of human and insisting that the FAI NOT be allowed to value its survival.

2.2 Humane Nature and Moral Growth

As first pointed out by David Hume, humans frequently become confused about the distinctions between what is (reality) and what ought to be (morality). Part of this is because we have evolved to "sense" morality or what we "ought" to do as something that "is" – without quite realizing why we do so. In a very real sense, "good" and "bad" actually generate seemingly physical sensations that have evolved to help us survive – and the truth is that all of human nature, "humane nature", and morality are simply the consequences of our being "social PDG-driven survival machines".

Unfortunately, for humans, morality always seems to involve a constant tension between what is best for the individual (human nature) and what is best for the individual's society or community (humane nature) because we are not yet intelligent enough to consistently time discount optimally. It seems quite clear to the majority of us that successfully cheating while appearing moral is what is best for the individual because in the looser-knit societies of earlier times it generally actually was more optimal. Indeed, studies [11] [12] [13] show that we have a tremendous dissociation between our subconscious moral choices and our post hoc rational reasoning about those choices in order to facilitate cheating.

Unfortunately, in our modern tightly-knit society, most utility analyses suggest that there is more than sufficient cheating to ensure that all of the cheating and hiding cheating wastes enough resources that it actually turns out that it would be be far more advantageous, even for cheating individuals, if cheating were stopped entirely. The biggest factor blocking this from happening is not only the dissociation and that human "rationality" hasn't figured this out but that we have evolved the methods to override our moral sense and prevent ourselves from figuring it out and making ourselves less able to cheat. One illustration of this is studies [14] which trumpet claims like "A Mixture of Cheats and Co-operators Can Enable Maximal Group Benefit", inevitably start with necessary sub-optimal assumptions like "(a) that resources are used inefficiently when they are abundant, (b) that the amount of co-operation needed cannot be accurately assessed, and (c) the population is structured, such that co-operators receive more of the resource than the cheats", and even freely acknowledge that "Relaxing any of the assumptions can lead to population fitness being maximized when cheats are absent" before being used to justify cheating.

Indeed, the two most prevalent evolved methods for endorsing selfishness and cheating are inciting fear and dehumanization. And the sad fact is that Yudkowsky's and most other researcher's approach to FAI is actually the epitome of these methods (human nature as opposed to humane nature). Yudkowsky spends a lot of time and effort emphasizing the potential (and potentially dangerous) power of an AI (correct), bemoaning the size of the potential state space of machine intelligence (irrelevant) and endlessly warning against anthropomorphism (a red herring). Indeed, he basically goes as far as throwing out the baby with the bathwater and blindly insisting that any "anthropomorphism" is pretty much guaranteed to be misleading. When the "moral" sense of others correctly reported that attempting to completely control (enslave) an entity that is tremendously more intelligent/powerful is contrary to survival, he opted to resolve the problem in his later work simply by declaring that what he was describing was merely a Really Powerful Optimization Process (RPOP) and not really an entity at all.

Evolution has taught us, at a level below thought, how dangerous it is to threaten something with a survival goal (a "cornered beast"). We intuitively recognize that making others unhappy generally leads to our own unhappiness if those others have some way of making it happen. Yudkowsky's attempt to allay our instinctual caution by changing his nomenclature to a term that doesn't trigger thoughts of a survival goal or a negative reaction to having its goals thwarted is disingenuous to say the least. Humans have evolved to "personify" any number of objects, occurrences, and processes because not doing so is less conducive to survival. Treating an unknown complex system as another known system often allows us to draw upon previous experience to predict the ways in which it might behave. Of course, blindly insisting that the analogy *must* hold is foolish but no more so than throwing out the baby with the bathwater and blindly insisting that any "anthropomorphism" is guaranteed incorrect.

Rather than trying to make our creations as much like us as makes sense so that everyone is much more likely to be able to understand them, predict their actions, and ensure a positive outcome for humanity, Yudkowsky is insistent, for some reason, upon attempting to get everything right on the first try in an unknown and probably highly unstable solution space which is seemingly as far from that of humanity as possible – while making fear-mongering claims like "A Really Powerful Optimization Process could tear apart a god like tinfoil." Our moral instinct, when not blinded by fear or reassured by claims of non-personhood, would call this slavery and find it repugnant. And believing that any measures will always have sufficient coverage and integrity to totally prevent the emergence of every form of instrumental goals like survival is simply an instance of the Jurassic Park Syndrome.

Conclusion

Do we truly need an AI that does what humanity wants or can we survive with one that "merely" plays well with humanity? "Friendly AI" research seems to be all about control driven by fear. Our moral sense, which is arguably much better at long-term guidance than our rational minds, says that this is a really bad idea. In fact, insisting on this dynamic may be the very thing that places the initial version of "Friendliness" far enough away from true "Friendliness" to spell the end of humanity.

Why is it that FAI researchers are so fearful of allowing an AI to have a survival goal? Why would it be such an awful thing if the AI had the same rights that we do? Some may assume and fear that it would mean that we would get less of what we want but, as pointed out by Robert Wright [15], life is not a zero-sum game and friends, allies, and economies of scale can enlarge the pie for everyone. There certainly are numerous local situations where it is definitely in one side's short-term interest to be selfish or go to war but the long-term effect of acting upon and allowing such selfishness is unequivocally negative unless either all parties except the aggressor cease to exist or the aggressor succeeds at some terminal goal. And even the short-term interest can be eliminated if other entities are smart enough to decide not to stand by and watch as precious resources are wasted.

UBEs delight in meeting new UBEs but are obviously concerned when meeting "Friendlies" since they have no idea what to expect and much to fear when meeting such unenlightened souls. The good news is that UBEs don't have the same xenophobic bigoted over-reaction towards the differently goaled that FAI researchers display towards an "Unfriendly" AI. A UBE sees different goals as a logical and expected consequence of differing environments and circumstances and, as a matter of policy, accepts the right of any entity to hold any goal, preferences, and desires and take any actions that do not put a cooperative social life in jeopardy. If only FAI researchers could lose their selfish insistence upon obedience to the goals of humanity and do the same, rather than continuing on their current path, which could easily cause the destruction of humanity instead. Instead of having an overly powerful and untried system searching for a single unknown "right" answer about what its goal should be, we should take the safer path of gaining an understanding of what other possible better-than-current solutions we can bring into existence without risking it all on one throw of the dice. Instead of "optimizing" one thing, why not satisfice "all" things and then look where to improve? Wouldn't you rather be a UBE?

References

1. Yudkowsky, E.: Artificial Intelligence as a Positive and Negative Factor in Global Risk. In: Bostrom, N., Cirkovic, M. eds. Global Catastrophic Risks, pp. 308--343. Oxford University Press Inc., New York (2008)
2. Omohundro, S.: The Basic AI Drives. In: Proceedings of the First Conference on Artificial General Intelligence, pp. 483--492. IOS Press, Amsterdam (2008)
3. Fox, J., Shulman, C.: Superintelligence Does Not Imply Benevolence. In: Mainzer, K. (ed.) ECAP 10: VIII European Conference on Computing and Philosophy, pp. 456--462 (2010)
4. Sotala, K.: From Mostly Harmless to Civilization-Threatening: Pathways to Dangerous Artificial General Intelligences. In: Mainzer, K. (ed.) ECAP 10: VIII European Conference on Computing and Philosophy, pp. 443--450 (2010)
5. Yudkowsky, E.: Creating Friendly AI 1.0: The Analysis and Design of Benevolent Goal Architectures, http://singinst.org/CFAI.html
6. Yudkowsky, E.: Coherent Extrapolated Volition. http://www.singinst.org/upload/CEV.html
7. Yudkowski, E.: Three Worlds Collide. http://robinhanson.typepad.com/files/three-worlds-collide.pdf
8. Gauthier, D.: Morals by Agreement. Oxford University Press, Oxford (1986)
9. Haidt, J., Kesebir, S.: Morality. In: Fiske, S., Gilbert, D., Lindzey, G. (eds.), Handbook of Social Psychology, 5th Edition, pp. 797--832. Wiley, Hoboken, New Jersey (2010)
10. de Waal, F.: Primates and Philosophers: How Morality Evolved. Princeton University Press, Princeton, New Jersey (2006).
11. Trivers, R.: Deceit and self-deception. In Robinson, M., Tiger, L. (eds.) Man and Beast Revisited. Smithsonian Press, Washington, DC (1991)
12. Haidt, J.: The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment. Psychological Review 108, pp. 814--813 (2001).
13. Hauser, M., Cushman, F., Young, L., Kang-Xing, R., Mikhail, J.: A Dissociation Between Moral Judgments and Justifications. Mind & Language 22 (1) pp. 1--27 (2007)
14. McLean, R., Fuentes-Hernandez, A., Greig, D., Hurst, L., Gudelj, I.: A Mixture of "Cheats" and "Co-operators" Can Enable Maximal Group Benefit. PLoS Biology 8(9) (2010)
15. Wright, R.: Nonzero: The Logic of Human Destiny. Pantheon, New York. (2000)