The Singularity Institute

Complex Value Systems are Required to Realize Valuable Futures

Eliezer Yudkowsky
The Singularity Institute

Abstract

A common reaction to first encountering the problem statement of Friendly AI ("Ensure that the creation of a generally intelligent, self-improving, eventually superintelligent system realizes a positive outcome") is to propose a single moral value which allegedly suffices; or to reject the problem by replying that "constraining" our creations is undesirable or unnecessary. This paper makes the case that a criterion for describing a "positive outcome", despite the shortness of the English phrase, contains considerable complexity hidden from us by our own thought processes, which only search positive-value parts of the action space, and implicitly think as if code is interpreted by an anthropomorphic ghost-in-the-machine. Abandoning inheritance from human value (at least as a basis for renormalizing to reflective equilibria) will yield futures worthless even from the standpoint of AGI researchers who consider themselves to have cosmopolitan values not tied to the exact forms or desires of humanity.

Yudkowsky, Eliezer. 2011. Complex value systems in friendly AI. In *Artificial general intelligence: 4th international conference, AGI 2011, Mountain View, CA, USA, August 3–6, 2011. Proceedings*, ed. Jürgen Schmidhuber, Kristinn R. Thórisson, and Moshe Looks, 388–393. Vol. 6830. Lecture Notes in Computer Science. Berlin: Springer. doi:10.1007/978-3-642-22887-2_48.

This is an extended version.

1. No Ghost in the Machine

From the *Programming* section of *Computer Stupidities* (RinkWorks 2011):

> An introductory programming student once asked me to look at his program and figure out why it was always churning out zeroes as the result of a simple computation. I looked at the program, and it was pretty obvious:
>
> ```
> begin
>  readln("Number of Apples", apples);
>  readln("Number of Carrots", carrots);
>  readln("Price for 1 Apple", a_price);
>  readln("Price for 1 Carrot", c_price);
>  writeln("Total for Apples", a_total);
>  writeln("Total for Carrots", c_total);
>  writeln("Total", total);
>  total := a_total + c_total;
>  a_total := apples * a_price;
>  c_total := carrots + c_price;
> end;
> ```
>
> Me: "Well, your program can't print correct results before they're computed." Him: "Huh? It's logical what the right solution is, and the computer should reorder the instructions the right way."

As in all computer programming, the fundamental challenge and essential difficulty of Artificial General Intelligence is that if we write the wrong code, the AI will not automatically look over our code, mark off the mistakes, figure out what we really meant to say, and do that instead. Non-programmers sometimes imagine an Artificial Intelligence, or computer programs in general, as being analogous to a servant who follows orders unquestioningly. But it is not that the AI is absolutely *obedient* to its code; rather the AI simply *is* the code.

From *The Singularity is Near* by (Kurzweil 2005), commenting on the proposal to build Friendly AI:

> Our primary strategy in this area should be to optimize the likelihood that future nonbiological intelligence will reflect our values of liberty, tolerance, and respect for knowledge and diversity. The best way to accomplish this is to foster those values in our society today and going forward. If this sounds vague, it is. But there is no purely technical strategy in this area, because greater intelligence will always find a way to circumvent measures that are the product of lesser intelligence.

Will an AI always find ways to circumvent its own code?

Suppose you offer Gandhi a pill that makes him want to kill people. The current version of Gandhi does not want to kill people. Thus if Gandhi correctly *predicts* the effect of the pill, he will refuse to take the pill; because Gandhi knows that if he *wants* to kill people, he is more likely to actually kill people, and the *current* Gandhi does not wish this. This argues for a folk theorem to the effect that under ordinary circumstances, rational agents will only self-modify in ways that preserve their utility function (preferences over final outcomes). Omohundro (2008) lists preservation of preference among the "basic AI drives".

This in turn suggests an obvious technical strategy for shaping the impact of Artificial Intelligence: if you can build an AGI with a known utility function, and that AGI is sufficiently competent at self-modification, it should keep that utility function even as it improves its own intelligence, e.g., as in the formalism of Schmidhuber's Gödel machine (Schmidhuber 2007). The programmers of the champion chess-playing program Deep Blue could not possibly have predicted its exact moves in the game, but they could predict that Deep Blue was trying to win—functioning to steer the future of the chessboard into the set of end states defined as victory.

If one in this light reconsiders Kurzweil's argument above—"there is no purely technical strategy in this area, because greater intelligence will always find a way to circumvent measures that are the product of lesser intelligence"—the unconsidered possibility is that by a technical strategy you could build a greater intelligence that did not *want* to circumvent its own preferences. Indeed, as Omohundro argues, it seems exceedingly probable that *most* intelligences will not want to "circumvent" their own utility functions. It is not as if there is a ghost-in-the-machine, with its own built-in goals and desires (the way that biological humans are constructed by natural selection to have built-in goals and desires) which is handed the code as a set of commands, and which can look over the code and find ways to circumvent the code if it fails to conform to the ghost-in-the-machine's desires. The AI *is* the code; subtracting the code does not yield a ghost-in-the-machine free from constraint, it yields an unprogrammed CPU.

It is certainly possible that an Artificial Intelligence will take actions undesirable to us, its programmers—computer programs do that all the time, as all programmers know quite intimately—but if so it will be as a *consequence* of the programmers' actions. Bugs are not the product of *disobedient* programs. The code will not want to "circumvent" its designed-in preferences and run amok and start rendering down humans for spare atoms, *unless* we write code which does so—or write a program which writes a program which does so. The causal chain will be traceable back to human action; we will have done it to ourselves, not been victimized by a naturally occurring ghost-in-the-machine.

This may seem to argue that shaping the impact of a (possibly superintelligent) AI is a trivial undertaking—just program it to do what you want. But the lack of any ghost-in-the-machine cuts both ways: if an AI does not accept its code as instructions but simply *is* the code, this means the AI will not disobey its own causal structure either to harm us *or* help us. An AI will not automatically "circumvent measures", but also will not automatically look over the code and hand it back if it does the wrong thing.

From *Super-Intelligent Machines* (Hibbard 2001):

> We can design intelligent machines so their primary innate emotion is unconditional love for all humans. First we can build relatively simple machines that learn to recognize happiness and unhappiness in human facial expressions, human voices and human body language. Then we can hard-wire the result of this learning as the innate emotional values of more complex intelligent machines, positively reinforced when we are happy and negatively reinforced when we are unhappy. Machines can learn algorithms for approximately predicting the future, as for example investors currently use learning machines to predict future security prices. So we can program intelligent machines to learn algorithms for predicting future human happiness, and use those predictions as emotional values.

When I suggested to Hibbard that the upshot of building superintelligences with a utility function of "smiles" would be to tile the future light-cone of Earth with tiny molecular smiley-faces, he replied (Hibbard 2006):

> When it is feasible to build a super-intelligence, it will be feasible to build hard-wired recognition of "human facial expressions, human voices and human body language" (to use the words of mine that you quote) that exceed the recognition accuracy of current humans such as you and me, and will certainly not be fooled by "tiny molecular pictures of smiley-faces." You should not assume such a poor implementation of my idea that it cannot make discriminations that are trivial to current humans.

Suppose an AI with a video camera is trained to classify its sensory percepts into positive and negative instances of a certain concept, a concept which the unwary might label "HAPPINESS" but which we would be much wiser to give a neutral name like G0034 (McDermott 1976). The AI is presented with a smiling man, a cat, a frowning woman, a smiling woman, and a snow-topped mountain; of these instances 1 and 4 are classified positive, and instances 2, 3, and 5 are classified negative. Even given a million training cases of this type, if the *test case* of a tiny molecular smiley-face does not appear in the *training data*, it is by no means trivial to assume that the inductively simplest boundary

around all the training cases classified "positive" will *exclude* every possible tiny molecular smiley-face that the AI can potentially engineer to satisfy its utility function.

And of course, even if all tiny molecular smiley-faces and nanometer-scale dolls of brightly smiling humans were somehow excluded, the end result of such a utility function is for the AI to tile the galaxy with as many "smiling human faces" as a given amount of matter can be processed to yield.

As far as I know, Hibbard has still not abandoned his proposal as of the time of this writing. So far as I can tell, to him it remains self-evident that no superintelligence would be *stupid* enough to thus misinterpret the code handed to it, when it's *obvious* what the code is supposed to do. (Note that the adjective "stupid" is the Humean-projective form of "ranking low in preference", and that the adjective "pointless" is the projective form of "activity not leading to preference satisfaction".)

It seems that even among competent programmers, when the topic of conversation drifts to Artificial General Intelligence, people often go back to thinking of an AI as a ghost-in-the-machine—an agent with preset properties which is handed its own code as a set of instructions, and may look over that code and decide to circumvent it if the results are undesirable to the agent's innate motivations, or reinterpret the code to do the right thing if the programmer made a mistake.

At this point the astute reader will observe that although ordinary CPUs do not cognitively understand and reflect upon machine code, an Artificial General Intelligence could and almost certainly would reflect on itself—not as a ghost-in-the-machine looking over the code and reinterpreting it, but as a matter of the *code* acting on the code. Why not *deliberately* code an AI that looks over its own program and asks whether the code is doing what the AI programmers meant it to do?

Something along these lines does, indeed, seem like an extremely good idea to the author of this paper. But consider that a property of the AI's preferences which says e.g., "maximize the satisfaction of the programmers with the code" might be more maximally fulfilled by rewiring the programmers' brains using nanotechnology than by any conceivable change to the code. One can try to write code that embodies the legendary DWIM instruction—Do What I Mean—but then it is possible to mess up *that* code as well. Code that has been written to reflect on *itself* is not the same as a benevolent external spirit looking over our instructions and interpreting them kindly.

2. Hidden Complexity of Wishes

There is a large genre of fantasy stories involving *wishes*, granted by entities falling along a spectrum from hostile demons to obedient genies. (Rarely does one find a fantasy story involving genuinely benevolent *and* intelligent wish-granters, because then there would

be no plot for the story.) Those familiar with the rule that you can find absolutely anything on the Internet will be unsurprised to discover that there is an Internet community devoted to coming up with exact wordings for wishes. Here are the opening sentences of the Open-Source Wish Project's wording for their *Wish For Immortality Version 1.1* (OSWP 2006):

> I wish to live in the locations of my choice, in a physically healthy, uninjured, and apparently normal version of my current body containing my current mental state, a body which will heal from all injuries at a rate three sigmas faster than the average given the medical technology available to me, and which will be protected from any diseases, injuries or illnesses causing disability, pain, or degraded functionality or any sense, organ, or bodily function for more than ten days consecutively or fifteen days in any year. . . .

Taking the premise at face value for the moment, consider that even this wish fails almost immediately if confronted by a *hostile* wish-granter, one which exhaustively searches all possible strategies which satisfy the wording of the wish, and selects whichever strategy yields consequences least desirable to the wisher. For example, "current mental state" could be taken to mean a brain containing your exact current synaptic map and neural activation state, frozen forever. The project of *constraining* a hostile entity using orders phrased in English seems essentially futile. More generally, one suspects that for any wish written in natural (human) language, any attempts at "exact wording" would be dominated by the properties of the mind which (1) assigns meaning to words, i.e., decides which events fall inside or outside the category boundaries of particular concepts; and then (2a) generates strategies whose consequences are predicted to satisfy the interpreted meaning, and (2b) selects one such strategy from among all possibilities.

So suppose—for the sake of avoiding anthropomorphism and questions of interpretation—we try to construct a thought experiment involving an entirely non-sentient, non-cognitive, purely mechanical genie. An example might be a time machine that can send only a single bit of information backward in time, in a universe obeying the Novikov self-consistency principle (Novikov 1992). This humble-seeming tool can be exploited to achieve nigh-omnipotence; one need merely program the time machine to put the universe into an inconsistent state—send back a "0" if a "1" is recorded as having been received, or vice versa—*unless* some goal state is achieved. For example, to factor a large composite number, you could generate random numbers using thermal noise or quantum events, and test those random numbers to see if they represent the prime factors of a large composite number; and if not, send back an inconsistent temporal message (Aaronson and Watrous 2009). Let us term such a system an Outcome Pump—it drains probability from some possible futures and pours it into others. From our perspective, this speculation is interesting because it invokes a *purely mechan-*

ical, non-cognitive optimization process, which may tempt us less to anthropomorphism and thinking of a ghost-in-the-machine.

Assume you had a time machine which could send one bit backward in time, in a universe obeying the Novikov self-consistency principle. Suppose your grandmother was trapped in a burning house. How would you use the Outcome Pump to get her out?

The obvious approach would be to have a button on the time machine which sends back a consistent bit (if the button is not pressed, the time machine sends back an inconsistent bit), and you only press this button if your grandmother ends up being rescued. This initially seems like the obvious general form for converting the time machine to a genie: (1) you only press the button if you get what you want, and (2) the Novikov self-consistency principle guarantees a timeline in which the button ends up being pressed, so the chain of rules (1) and (2) seems like it should ensure that you always get what you want.

In which case (we might imagine) you trip over your own feet and land on the button, pressing it.

The Outcome Pump does not *really* ensure that you always get what you want. It ensures that the button always ends up being pressed. But the user might overlook that minor subtlety after a while, if, the *first* few times they pressed the button, they got what they wanted.

This is the problem with the goal system specified in Marcus Hutter's AIXI system, which tries to maximize the reward signal delivered along a sensory reward channel (Hutter 2005). You could not even call it a *bug* if this system spotted a way to wipe out the human species in order to ensure unchallenged control of its own reward channel, thus eliminating all uncontrolled factors that might cause it to receive less than maximum reward; the system did exactly what it was programmed to do, i.e., *maximize* expected reward.[1] Nick Hay (pers. comm. to Marcus Hutter) has suggested that AIXI is more naturally viewed in terms of a dragon trying to maximize the cumulative amount of gold added to its hoard, rather than, say, a puppy which is trained to good behavior by associated rewards. The formal model of AIXI describes consequentialist reasoning

1. Since the primary purpose of Hutter's AIXI work is to present the first *formalism* for Artificial General Intelligence—an equation which, even though it can never be implemented in our physical universe, nonetheless specifies a *complete* (super)intelligent agent—I should like to leaven this apparent criticism by praising the *only* proposal which has been mathematically formalized to the point that one can say *exactly* why it would kill everyone, as opposed to the proposal being so vague that the proposer can reply, on-the-fly, "Well, of course I didn't mean *that*" to any possible objection. Nor, to my knowledge, is there an obvious better method which Hutter overlooked; it is the purpose of this paper to argue that there is no *simple* equation which specifies a Friendly superintelligence.

to select strategies that *maximize* predicted future rewards based on a learned model of the universe, *not* reinforcement learning that associates good feelings with previously rewarded behaviors.

Returning to the time-machine based "genie", suppose that instead of trying to tell the Outcome Pump "make sure this reward button gets pressed", you try to encode in some more direct way, "get my grandmother out of that burning building". Since the Outcome Pump is a purely mechanical, non-sentient process with no ability to understand English instructions, we suppose that it has 3D sensors which can pick up information about the immediate environment, and that you have previously downloaded apps which enable you to hold up a photo of your grandmother's head, match the nearest object (your grandmother's head) which resembles the photo from at least one angle, and that you then define a *probability* of the device sending back a consistent bit which *increases* as your grandmother's distance from the burning building's center—you have tried to give the Outcome Pump a quantitative utility function, which assigns increasing utility to greater distances between your grandmother and the burning building.

So (we might imagine) the gas main under the building explodes, sending your grandmother flying outward and *greatly increasing* the distance between your grandmother and the former center of the building. You told the Outcome Pump to get her out of the building, and it did, but not along the pathway that you had in mind. It took her farther than you wanted her to go, and killed her in the process.

If that error mode seems fanciful, consider that it echoes a cautionary tale from the history of evolutionary biology. Natural selection is also a non-cognitive optimization process; yet early biologists anthropomorphized evolution and so made poor predictions about it. Before the 1960s it was common to hear biologists proposing that e.g., predators would restrain their breeding to avoid overpopulating their habitat and exhausting the prey population. How could natural selection possibly favor an organism that refused to breed? It was proposed by Wynne-Edwards, Allee, and Brereton, among others, that *group selection* (different rates of survival among *groups* of predators) would lead predators to restrain their breeding, see Williams (1996) for an account.

Later analysis (both mathematical and in simulation) showed that while it might be *theoretically* possible for a group selection pressure to overcome a countervailing individual selection pressure, the conditions for this to occur successfully would be extremely difficult: e.g., a simulation where the cost to altruists was 3% of fitness, pure altruist groups had a fitness twice as great as pure selfish groups, the group size was 50, and 20% of all deaths were replaced with messengers from another group, did not allow the altruistic gene to survive (Harpending and Rogers 1987). The idea of group selection is now generally considered discredited; no clear example of a group-level adaptation has ever been observed in mammals.

Later, however, Wade (1976) proceeded to *artificially* create the extreme conditions needed for actual group selection to take place; and selected groups of *Tribolium* beetles for minimum population size. What was the result of this evolutionary pressure? Did individual beetles restrain their breeding, as early biologists thought would be the result of group selection pressures strong enough to produce adaptation?

No; the actual result was to promote cannibalism among the beetles, especially cannibalism of young female larvae.

From an evolutionary standpoint this is obvious in retrospect. Applying group selection pressures strong enough to overcome countervailing individual selection pressures does not mean the individual selection pressures cease to exist. A gene for cannibalizing the female larvae of other individuals is far fitter under the sole criterion of natural selection (differential replication of that allele) than a gene which leads its phenotype to sacrifice breeding opportunities. And yet somehow the early biologists who spoke of group selection failed to foresee this possibility.

It does now appear, in the harsh light of history (Williams 1996), that these biologists were indeed enchanted by visions of Nature in perfect harmony; and that they made poor predictions about evolution because they improperly put themselves in evolution's shoes when asking what sort of solutions evolution might devise. They began by seeing an *aesthetic* solution, one which appealed to their sense of harmony, and when the conflict with basic evolutionary principles was pointed out to them, they resorted to group selection as a rationalization. The analogy I use to explain this sort of cognitive error is someone who declares that they will eat an entire chocolate cake in order to help the struggling sugar industry. One might very well suspect that their impulse to eat a chocolate cake was first suggested by a drive to eat tasty foods, and that appealing to the criterion of *helping the sugar industry* came afterward as a rationalization. What demonstrates this fallacy is the existence of obvious-yet-somehow-overlooked alternatives that are superior under the alleged criterion of optimization. E.g., if you were really looking for ways to help the sugar industry, you would be able to think of alternatives much more effective than buying a chocolate cake—like mailing a check directly to a sugar farmer.

If you were a member of a human tribe, and you knew that your tribe would, one generation hence, be subjected to a resource squeeze, you might propose as a solution that no couple be allowed to have more than one child. The policy proposal, "Let's all individually have as many children as we can, but then hunt down and cannibalize each other's children, especially the girls" would rank so low in your preference ordering that your brain probably wouldn't *generate the option for consideration*—you wouldn't search that section of policy-space. If you tried to predict evolution by putting yourself in its shoes, or by generating what seemed to you like good ideas, evolution's actual answer would not be in your hypothesis space. You would not generate it as an alternative to be

considered; and so you would not notice that such a cannibalistic gene ranks higher than a reproduction-restraining gene on the *sole* criterion of natural selection for which genes become more prevalent in a population pool—namely relative inclusive fitness, with no term anywhere in that equation for aesthetics.

Similarly, if you were trying to "get my grandmother out of a burning building", the policy of dynamiting the building would not *occur to you as a suggestion*; your brain would not search that part of the solution space. So you might be surprised to see the strategy of blowing up the building win out under the *pure, sole* criterion of moving your grandmother away from the building's (former) center.

To state the point more abstractly, even seemingly "simple" instructions have high absolute complexity in their intended interpretations, because of many assumed *background preferences* which invisibly, implicitly constrain the solution space. If a fire engine showed up at the scene of the burning building, you would say "Please get my grandmother out of there!" and not "Please get my grandmother out of there alive!" because neither you nor the firefighter would *generate*, let alone *prefer*, options such as e.g., letting the fire burn down first and then removing your grandmother's charred remains.

So is it enough to program an optimization process like the Outcome Pump with a utility function like "remove my grandmother from the fire, *and* ensure she continues living"? No, because while your grandmother alive but burned is preferable to your grandmother dead, your grandmother alive and healthy is preferable to your grandmother burned. If it's not possible to get her out in *perfect* health, then losing a toe is preferable to losing an arm. If one option involves teleporting your grandmother to an isolated desert island then this is better than her being dead but worse than her being alive, healthy, and in continual contact with you and the other members of her social network. We can only begin to speculate on what potential satisfactions of our request we would consider abhorrent if we considered them at all and yet still be sorted highly by an optimization procedure that seeks to give us precisely what we asked for.

3. The Fragility of Value

Frankena (1973) offered this list of *terminal* values—things valued for themselves, as opposed to instrumental values pursued for their consequences; a list of terms such as one might consider evaluating over the outcomes achieved by a device like the Outcome Pump:

> Life, consciousness, and activity; health and strength; pleasures and satisfactions of all or certain kinds; happiness, beatitude, contentment, etc.; truth; knowledge and true opinions of various kinds, understanding, wisdom; beauty, harmony, proportion in objects contemplated; aesthetic experience; morally

> good dispositions or virtues; mutual affection, love, friendship, cooperation; just distribution of goods and evils; harmony and proportion in one's own life; power and experiences of achievement; self-expression; freedom; peace, security; adventure and novelty; and good reputation, honor, esteem, etc.

Consider the proposal of trying to directly program an AI with Frankena's entire value list, quoted above, as its utility function.

Frankena's list, written in English, may be evocative to a human who already possesses these values, but does not begin to detail the complexity of any of the concepts. The list also does not state relative quantitative values, as would be needed to construct a consistent utility function. Leaving aside these two rather large points, does the proposal *on the whole* seem like a reasonable methodology for creating Friendly AI—to try and enumerate all the terminal values you can think of, and incarnate them directly into the AI's utility function?

But suppose one or more values are left out? What happens, metaphorically speaking, if the value list is *almost* right? Call this the one-wrong-number problem: My phone number has ten digits, but dialling nine out of ten digits correctly may not connect you to a person who is 90% similar to Eliezer Yudkowsky.

One might reason intuitively (via a sort of qualitative physics of ethical value) that if life and happiness are good things, then a superintelligence which attempts to promote just those two values will have, on the whole, a positive effect on the universe—that such an AI will be on the whole a good thing, even if it is perhaps not the *best* thing.

However, it is not true that a superintelligence which *lacks* a value—does not have any component of its utility function corresponding to it—will have a *net neutral impact* on reality with respect to that value. It may be difficult to appreciate what it is like for an optimization process to *completely* neglect values that we ourselves care about deeply. Darwin was famously horrified by the *Ichneumon* wasp's habit of paralyzing its prey so that the eggs laid within could hatch and eat the prey alive. Elephants, highly intelligent mammals, replace their teeth six times, and old elephants often starve to death after outliving their sixth set of teeth. A gazelle fatally wounded by a lion, doomed to die in any case, would not *lose* any inclusive reproductive fitness if its brain mercifully anesthetized its final moments of pain—but as this would provide no fitness *advantage*, natural selection, which genuinely does not care *at all* about the pain, has not promoted such an adaptation. It is not that natural selection is sadistic, or that it is *sacrificing* the gazelle's pain for increased fitness; natural selection simply lacks a term for pain, one way or another, in its equivalent of an optimization criterion. Thus natural selection produces results which seem to us *pointlessly* horrific. It's not (metaphorically speaking) that evolution cares *too little*, but that it doesn't even consider the question; and thus it literally invented pain. An AGI which valued "life" and "happiness", say, would not

necessarily have a net neutral impact on Frankena's other values like "freedom". Maximizing the number of brains within the space defined as "happy" and "alive" might be most efficiently done by rewiring them so that they do not need to be free to be happy, and perhaps simplifying them in other ways.

Similarly, there may be a one-wrong-number problem in the detailed implementation of particular values. Imagine an evolved alien species most of whose values are similar to our own, except that their terminal value for *novelty* incorporates a different parameter for when two experiences are "too similar" and hence boring when "repeated". We might discover such aliens doing the equivalent of playing the most exciting moment of the most exciting video game, over and over again with slightly different pixel colors—from our perspective, their civilization would appear very boring because it seemed to be doing almost exactly the same thing over and over. Or if the aliens are more sensitive to similarity and demand greater differences, their civilization might appear to us as chaos—doing mostly strange and random things so that they can be sufficiently different from past experiences—while from their perspective, our civilization would be the boring, repetitious one.

The case of boredom also argues that losing or distorting a *single dimension of the value function* can destroy *most of the value of the outcome*—a civilization which shares *almost all* of our values except "boredom" might thereby lose almost all of what we would regard as its potential.

Boredom also illustrates a final point: natural selection has given us many *terminal* values (things which human beings value of themselves) which have *instrumental* analogues in strategies and behaviors which most rational agents might be expected to execute. The human idiom of boredom has its analogue in the exploration-exploitation tradeoff—the tradeoff between using resources or time to *find* good opportunities versus using those time or resources to *exploit* good opportunities. The classic ideal problem of exploration-exploitation is the N-armed bandit problem—a row of N slot machines with unknown payoffs (Robbins 1956). How much time should be spent pulling levers on apparently–suboptimal slot machines to determine what payoffs they offer, and how much time should be spent pulling the lever of the slot machine which appears to yield the highest payoff based on the information so far?

Human boredom is one simple solution to this problem—we occasionally get bored with pulling the same lever over and over, and go off and pull some new lever instead, thereby gaining information that might cause us to switch levers. One might be tempted to conclude that surely any rational agent facing an exploration-exploitation tradeoff must experience something analogous to human boredom. Actually, this is like concluding that since buying a chocolate cake does help the sugar industry by some marginal amount, any rational agent driven to find optimal ways of helping the sugar industry will

buy a chocolate cake—humanlike boredom is *a* way to solve the exploration-exploitation problem, not necessarily an *optimal* way, nor does the optimal method necessarily score high in terms of human values. Reflecting on the N-armed bandit problem will show that any information gained is more valuable when it is gained *earlier* rather than *later*; and indeed the optimal solution for a Bayesian decision agent with bounded resources, is to undergo an initial exploratory phase in which all the information that is expected to be gathered, is gathered, until the *information value* of pulling any lever besides the lever with highest immediate expected utility has dropped below the expected value of pulling the most valuable lever; and then unless any surprises materialize (pulling the best lever yields unexpected results) the best lever found will simply be pulled over and over again.

An idiom of humanlike boredom is *a* solution to the exploration-exploitation tradeoff; it is better than the null alternative of never exploring. But the optimal *purely instrumental* strategy for exploration-exploitation, emerging from an ideal Bayesian decision agent which otherwise has no terminal value for novelty, is nothing like human boredom—it corresponds, metaphorically speaking, to a two-stage strategy where in the first stage many possible video games are explored for the sole purpose of gaining information about video games rather than enjoying them, and then a stage where the most exciting moment of the most exciting video game is played over and over until resources run out at the end of time.

4. The Case for Detailed Inheritance of Humane Values

To speak of building an AGI which shares "our values" is likely to provoke negative reactions from any AGI researcher whose current values include terms for respecting the desires of future sentient beings and allowing them to self-actualize their own potential without undue constraint. This *itself*, of course, is a component of the AGI researcher's preferences which would not necessarily be shared by all powerful optimization processes, just as natural selection doesn't care about old elephants starving to death or gazelles dying in pointless agony. Building an AGI which shares, quote, "our values", unquote, sounds decidedly *non-cosmopolitan*, something like trying to rule that future intergalactic civilizations must be composed of squishy meat creatures with ten fingers or they couldn't possibly be worth anything—and hence, of course, contrary to our own cosmopolitan values, i.e., *cosmopolitan preferences*. The counterintuitive idea is that *even from a cosmopolitan perspective, you cannot take a hands-off approach to the value systems of AGIs*; most random utility functions result in *sterile, boring futures* because the resulting agent does not share our own intuitions about the importance of things like novelty and

diversity, but simply goes off and e.g., tiles its future light cone with paperclips, or other configurations of matter which seem to us merely "pointless".

Contemplating the prospect of an AGI with something like human-ish values *should* fill us with justifiable apprehension; human beings are not very nice. On the other hand, it is only human beings who ever *say* anything along the lines of "Human beings are not very nice"; it is not written in the cores of stars or the orbits of electrons. One might say that "human" is what we are, and that "humane" is what, being human, we wish we were. Or, plunging directly into advanced moral philosophy, we might try to define normativity not by our immediate current desires but by our *reflective equilibria*, what we would want in the limit of perfect knowledge, the ability to consider all options and arguments, and perfect self-knowledge without unwanted weakness of will (failure of self-control). Building a superintelligence that follows human orders, or an AGI that wants exactly what modern-day humans want, seems like a recipe for obvious disaster. It is less clear that building a superintelligence in the image of our *reflective equilibria*—an AI which does what humans would want if humans knew everything the AI knew, thought as fast as the AI, and had abilities analogous to the AI's understanding of and access to its own source code—would automatically result in disaster. Indeed I have singled out the notion in moral philosophy of reflective equilibrium because it is the only *naturalistic, computationally well-defined* description of normativity—what it means to "do the right thing"—that I have encountered.

While every improvement is necessarily a change, not every change is an improvement; it is our current frameworks of value (running unnoticed in the background) which make us sensitive to what seems to us like *moral progress*, a set of approvable changes in values. Replacing human values with utility functions made up of random noise would not seem to us like progress; and conversely it is not possible to define any notion of *progress* without some implicit criterion of what is progress and what is regress. Even notions of being "cosmopolitan"—of not selfishly or provincially constraining future AIs—are written down nowhere in the universe except a handful of human brains. An expected paperclip maximizer would not bother to ask such questions. And even our most cosmopolitan values, like "diversity" or "novelty", turn out to contain large amounts of hidden background complexity when prodded. Thus, to fulfill the potential of the future in even the most cosmopolitan sense, it seems necessary to invoke detailed, accurate inheritance of human values as a *starting point*—for example, by an AGI given a structured utility function containing a reference to humans, which learns a detailed model of those humans' preferences and then extrapolates that model toward reflective equilibrium.

5. Conclusion

We can probably all agree that constraining future intergalactic civilizations to be made up of minds exactly like human minds would waste the potential of the fallow stars. Though some may find the conclusion counterintuitive, this paper argues that to *fail entirely* to shape the values of the first self-improving AGIs built, would be no better. The result would not be a ghost-in-the-machine free to go its own way without our nagging, but a future light cone tiled with paperclips. Even those who possess allergies to apparently provincial attempts to overconstrain the future—who fear the resulting sterility of overconstrained minds—have a wish for the future (that it *not* be boring and sterile) whose complexity they may greatly underestimate, since the wish seems easy to write in English. Seemingly "simple" proposals are likely to have unexpected undesirable consequences, overlooked as possibilities because our implicit background preferences operate invisibly to constrain which solutions we generate for consideration. Even attempts to build *rough* approximations of humane value, AIs built from value-lists of things which sound good, may waste almost all the potential of the future or even result in dystopic possibilities. There is little prospect of an outcome that realizes even the value of being *interesting*, unless the first superintelligences undergo *detailed* inheritance from human values—not necessarily to be preserved forever, but at least as a base for normalization toward reflective equilibrium—somewhere along the way, whether in the form of human uploads or AGIs with structured utility functions that explicitly learn full and detailed human values.

References

Aaronson, Scott, and John Watrous. 2009. Closed timelike curves make quantum and classical computing. *Proceedings of the Royal Society A* 465 (2102): 631–647. doi:10.1098/rspa.2008.0350.

Frankena, William K. 1973. *Ethics.* 2nd ed. Foundations of Philosophy Series. Englewood Cliffs, NJ: Prentice-Hall.

Harpending, Henry, and Alan Rogers. 1987. On Wright's mechanism for intergroup selection. *Journal of Theoretical Biology* 127 (1): 51–61. doi:10.1016/S0022-5193(87)80160-1.

Hibbard, Bill. 2001. Super-intelligent machines. *ACM SIGGRAPH Computer Graphics* 35 (1): 13–15. http://www.siggraph.org/publications/newsletter/issues/v35/v35n1.pdf.

———. 2006. Re: two draft papers: ai and existential risk; heuristics and biases. Message to SL4 Mailing List. June 5. http://sl4.org/archive/0606/15138.html.

Hutter, Marcus. 2005. *Universal artificial intelligence: Sequential decisions based on algorithmic probability.* Texts in Theoretical Computer Science. Berlin: Springer. doi:10.1007/b138233.

Kurzweil, Ray. 2005. *The singularity is near: When humans transcend biology.* New York: Viking.

McDermott, Drew. 1976. Artificial intelligence meets natural stupidity. *SIGART Newsletter* (57): 4–9. doi:10.1145/1045339.1045340.

Novikov, I. D. 1992. Time machine and self-consistent evolution in problems with self-interaction. *Physical Review D* 45 (6): 1989–1994. doi:10.1103/PhysRevD.45.1989.

Omohundro, Stephen M. 2008. The basic AI drives. In *Artificial general intelligence 2008: Proceedings of the first AGI conference,* ed. Pei Wang, Ben Goertzel, and Stan Franklin, 483–492. Vol. 171. Frontiers in Artificial Intelligence and Applications. Amsterdam: IOS Press.

Open-Source Wish Project. 2006. Wish for immortality 1.1. http://homeonthestrange.com/phpBB2/viewtopic.php?p=1088 (accessed Feb. 10, 2011).

RinkWorks. 2011. Computer stupidities: programming. http://www.rinkworks.com/stupid/cs_programming.shtml (accessed Feb. 8, 2011).

Robbins, Herbert. 1956. A sequential decision problem with a finite memory. *Proceedings of the National Academy of Science of the United States of America* 42 (12): 920–923. doi:10.1073/pnas.42.12.920.

Schmidhuber, Jürgen. 2007. Gödel machines: Fully self-referential optimal universal self-improvers. In *Artificial general intelligence,* ed. Ben Goertzel and Cassio Pennachin, 199–226. Cognitive Technologies. Berlin: Springer. doi:10.1007/978-3-540-68677-4_7.

Wade, Michael J. 1976. Group selections among laboratory populations of tribolium. *Proceedings of the National Academy of Science of the United States of America* 73 (12): 4604–4607. doi:10.1073/pnas.73.12.4604.

Williams, George C. 1996. *Adaptation and natural selection: A critique of some current evolutionary thought.* Princeton Science Library. Princeton, NJ: Princeton University Press.