Superintelligence as superethical

Steve Petersen*

2016-09-14

Human extinction by evil, super-intelligent robots is standard fare for outlandish science fiction—but Nick Bostrom's book *Superintelligence* (2014) summarizes a robot apocalypse scenario worth taking very seriously. The story runs basically like this: once we have a machine with genuine intelligence like ours, it will quickly be able to design even smarter and more efficient versions, and these will be able to design still smarter ones, until AI explodes into a "superintelligence" that will dwarf our own cognitive abilities the way our own abilities dwarf those of a mouse.[1] There is no special reason to think this superintelligence will share any of our own goals and values, since its intelligence won't have been shaped by the evolutionary history that endowed us with our particularly human needs (such as for companionship or salty snacks). Its ultimate goal might be simply to maximize the total number of paperclips in the world, if some enterprising paperclip company happens to be the first to stumble on the trick to genuine AI. Such a superintelligent Paperclip Maximizer, driven by its own internal value system rather than any actual malice, will quickly think of devastatingly effective ways to turn all available resources—including us humans—into paperclips.[2] All this could happen so fast that we wouldn't even have time for the luxury of worrying about any other ethical implications of genuine AI.[3]

Bostrom's concern is getting serious attention. For example, Stephen Hawking and a panoply of AI luminaries have all signed an open letter calling for more research into making AI safe, and entrepreneur Elon Musk has founded a billion-dollar nonprofit organization dedicated to this goal.[4] Such portents may seem overly dramatic, but it's worth remembering that it only takes one big event to wipe us out, and the fact that we've so far survived other risks (such as a nuclear war or pandemic) is no evidence that we *tend* to survive them—since we couldn't be around to observe the risks we *don't* survive.[5] Existential risks

*Thanks to Rob Bensinger, John Keller, Robert Selkowitz, and Joe Stevens.

[1]This idea of an intelligence explosion, or "singularity", resulting from AI goes back to another unsung statistics hero from Bletchley Park: Jack Good (1965). The mouse analogy is from Chalmers (2010).

[2]The "paperclip maximizer" example is originally from Bostrom (2003b).

[3]Implications such as whether genuinely intelligent robots could or should ethically be made to serve us, ahem—see Petersen (2012).

[4]See http://futureoflife.org/ai-open-letter/ and https://openai.com/about/.

[5]As a friend once put this point: "*Leap and the net will appear* was clearly not written

can boggle the mind, giving wishful thinking a chance to creep in where we need cold rationality. Bostrom warned in an interview that

> [p]eople tend to fall into two camps. On one hand, there are those ... who think it is probably hopeless. The other camp thinks it is easy enough that it will be solved automatically. And both of these have in common the implication that we don't have to make any effort now.[6]

I agree with Bostrom that the problem merits serious attention now. It's worth remembering, though, that resources spent on safe AI have real opportunity costs. Based on this risk assessment, philanthropists concerned to provide evidence-based, "effective altruism" are now diverting money to safe AI that otherwise would have gone toward saving people from starvation today.[7] And we must also factor in the added costs if excessive caution delays what Eliezer Yudkowsky (2008) calls a *friendly* superintelligence—especially one motivated to end famine, cancer, global warming, and so on.

So although care is certainly warranted, it's worth calibrating the risk level carefully, and that is why I propose to play devil's advocate with Bostrom's distressing argument. Appealing to a few principles that Bostrom already accepts, I argue here that ethical superintelligence is more probable than he allows. In summary, the idea is that a superintelligence cannot be prewired with a final goal of any real complexity, and so (Bostrom agrees) it must *learn* what its final goals are. But learning final goals is tantamount to *reasoning* about final goals, and this is where ethics can get a foothold.

Superintelligence and complex goals

In the positive portion of his book, Bostrom considers prospects for friendly AI. We would like to program the superintelligence to share goals like ours—but as Bostrom dryly notes, "human goal representations are complex" (p. 227), and so "explicitly coding [their] requisite complete goal representation appears to be hopelessly out of reach" (p. 228).

> Computer languages do not contain terms such as "happiness" as primitives. If such a term is to be used, it must first be defined. ... The definition must bottom out in terms that appear in the AI's programming language, and ultimately in primitives such as mathematical operators and addresses pointing to the contents of individual memory registers. (p. 227)

by someone who took a freefall. Those people are never heard from again." (Few if any know more about observation selection effects than Bostrom himself.)

[6] Khatchadourian (2015).

[7] See Matthews (2015).

Philosophers do not even agree on how to paraphrase *justice* or *happiness* into other similarly abstract terms, let alone into concrete computational primitives.

But human goals are hardly unique for being complex. Even the goal to "maximize paperclips" would be very difficult to program explicitly, and is radically underspecified as it stands. This worry is implicit in Bostrom's "perverse instantiation" cases (p.146), where superintelligences find literally correct but unintended ways to fulfill their goals—the way genies in fairy tales often fulfill wishes.[8] To give a taste for how the goal of "maximize paperclips" is underspecified: do *staples* count as paperclips? Do C-clamps count? Do they still count if they are only 20 nanometers long, and so unable to clip anything that would reasonably count as paper? Do they still count if they are so flimsy they would instantly snap should anyone attempt to use them? Do they count in structurally identical computer-simulated worlds? These questions may sound abstruse, but they matter when a superintelligent Paperclip Maximizer ("PM" for short) is trying to make the most possible paperclips. More to our point: do they still count as paperclips if they are just like the ones on our desks today, but they could never actually be used to clip paper (because any paper and any people to clip it are busy being turned into more "paperclips")? If paperclips must have a fighting chance of being *useful* to count, the PM will be considerably less threatening.

We could presumably program some of these answers in ahead of time, but there will still be plenty more leftover. Even providing a prototype to be scanned and saying "maximize things like this" requires specifying what it is to be "like" that. (Like that paperclip in terms of its history? In terms of the dust particles on its surface?) The point is that pursuing goals that are even a bit abstract requires too many fine-grained details to be programmed ahead.

So if we cannot wire complex goals ahead of time, how could the superintelligence ever possess them? Bostrom's various proposals, in the case of giving a superintelligence complex human values, all come down to this: the superintelligence must *learn* its goals.[9]

For an AI to learn a goal is not at all odd when it comes to *instrumental* goals—goals that themselves aim toward some further goal. Thus for example the PM might have an instrumental goal to mine a lot of ore. The PM is interested in mining only insofar as mining helps with its further goal of obtaining raw materials. Instrumental goals are just means to an agent's true end, and part of the whole point of AI is to devise new means that elude us. Indeed, a range of adaptability in ways to achieve an end is basically what folks in the AI community *mean* by "intelligence."[10] Instrumental goals are comparatively easy to learn,

[8] For example, if the final goal is to "maximize smiles", then the superintelligence could "tile the future light-cone of Earth with tiny molecular smiley-faces", as Yudkowsky (2011) points out. (This paper also has a nice comparison to the art of genie wishing.) If the final goal is to "make *us* smile", Bostrom points out the superintelligence could just "paralyze human facial musculatures" (p. 146).

[9] Though only Bostrom's favored approach actually has the word 'learning' it its name, they are all learning techniques in the more traditional AI sense.

[10] Bostrom says what he means by the word is "something like skill at prediction, planning,

since they have a clear criterion for success: if achieving that instrumental goal helps its further goal, keep it; if not, chuck it and try some other. This regress ends in a *final* goal—a goal sought for its own sake. The PM, for example, just seeks to maximize paperclips. Ultimately final goals serve as the learning standard for instrumental goals.

But Bostrom proposes the superintelligence learn its *final* goal, and that is a trickier matter. If the PM adjusts its final paperclip goal for a reason, it seems that means there must be some background standard the paperclip goal fails to achieve by the PM's own lights—which seems to mean that other background standard was its true final goal all along. On the other hand, if the PM has no deeper reason to change its final goal, then that goal change was arbitrary, and not learned. In general it seems learning requires a standard of correctness, but any standard against which a putatively final goal could be learned makes that further standard the *real* final goal. Thus it seems impossible to learn final goals. Call this simple argument *Hume's dilemma*, since it motivates David Hume's thesis that beliefs—even ethical ones—cannot influence goals without some other background goal (such as to be ethical) already in place.[11]

So it seems we can neither program a superintelligence's complex final goal ahead of time, nor have it learn the complex final goal on its own. It is telling that frontrunners for general AI, such as Marcus Hutter's AIXI and Karl Friston's free energy approach, simply take goal specification for granted in one way or another.[12]

And yet, learning new final goals seems like something we humans routinely do; we spend much of our lives figuring out what it is we "really" want. Furthermore, it feels like this is something we can make progress on—that is, we are not merely arbitrarily switching from one goal to another, but gaining *better* final goals. When Ebeneezer Scrooge comes to value warm cheer over cold cash, we think both that he has changed fundamental values, and that he is the better for it. Of course, we could just say that Scrooge always had the final goal of *happiness*, and that he has learned better instrumental means to this goal. But such a vague goal is unhelpful; as Aristotle noted thousands of years ago, "to say that happiness is the chief good seems a platitude, and a clearer account of what it is still desired."[13] It seems there is no sharp line between determining what one's final ends really are, on the one hand, and determining specific means to a vaguer but fixed final end on the other.

---

and means-ends reasoning in general" (p. 130). He is not alone in this usage; see for example Lycan (1987) p. 123, Clark (2001) p. 134, or Daniel Dennett's "Tower of Generate and Test" in *e.g.* Dennett (1994).

[11] Hume (1739) 2.3.3.

[12] See *e.g.* Hutter (2005) and Friston and Stephan (2007).

[13] Aristotle (Circa BCE 350), 1097b22.

Complex goals and coherence

The ethical view known as *specificationism* addresses this point. It holds that "at least some practical reasoning consists in filling in overly abstract ends ... to arrive at richer and more concretely specified versions of those ends."[14] Specificationism suggests there is no clear distinction between determining what one's final ends really are, on the one hand, and determining specific means to a more vague, but fixed, final end on the other. Specificationism responds to Hume's dilemma by suggesting that a final goal can be learned (or, if you like, *specified*) against a standard substantive enough to influence reasoning, but too formal to count as a goal itself—namely, the standard of overall *coherence*. The exact nature of coherence reasoning is itself up for grabs,[15] but the basic idea is to systematize a set of thoughts between which exist varying degrees of support and tension, without holding any special subgroup of thoughts as paramount or inviolable.

In the case of practical reasoning—reasoning about what to do—coherence must be found among potential goal specifications, potential routes to their success, and whatever other information might be relevant; roughly speaking, the coherence must be between beliefs about how the world is, and desires about how the world should be. A simple example of practical incoherence is a final goal specification that simultaneously demands *and* prohibits paperclips under 20nm in length. Such an incoherence must be reconciled somehow by appealing to tiebreakers. Similarly, if the PM believes there is no such thing as phlebotinum, then coherence prohibits a goal of making paperclips from the stuff. In this way beliefs can inform goal specifications. And conversely, its goal specifications will help it decide which truths to seek out of the impossibly many truths available, and so inform its beliefs; if the PM thinks that paperclips made of stronger, lighter material might best aid paperclip maximizing, then its goal would motivate it to study more materials science.

Bostrom proposes that an AI learn sophisticated goals using a value-learning model he calls "AI-VL", based on Dewey (2011). AI-VL is basically a coherence reasoning system. Ideally we would guide the superintelligence's actions by programming an exact value score for every possible set of circumstances—a "utility function". But since explicit utility functions are impossible for all but the very simplest of goals, the AI-VL model instead constructs an average utility function out of its weighted *guess*, for each possible utility function, that it is the *right* utility function (given the world in question) according to a "value

[14] Millgram (2008), p. 744. Key specificationist papers are Kolnai (1962) and Wiggins (1975).

[15] As Millgram (2008) puts it, "coherence is a vague concept; we should expect it to require specification; indeed, there are already a number of substantively different and less woolly variations on it, with indefinitely many more waiting in the wings" (p. 741). Thagard and Verbeurgt (1998) and Thagard (1988) are good places to start. In collaboration with Millgram, Thagard developed accounts of *deliberative* coherence in Millgram and Thagard (1996) and Thagard and Millgram (1995); see also Thagard (2000). Though inspired by such work, I now lean toward an alternative Millgram also mentions—see *e.g.* Grünwald (2007).

criterion". Now this is not anything like a ready-to-go solution. Besides being "wildly computationally intractable" (p. 239), this approach pushes most of the problem back a step: it is a mystery how we could specify a detailed value criterion in a way largely under our control, and a mystery how its probabilities might be updated. But it is an interesting proposal, and supposing we could get it to work, the important point for our purposes is that such a superintelligence would be using its beliefs about the world (its guesses about the right utility function) to figure out (or specify) what its final goals are, while simultaneously using its goals to figure out what beliefs to form. In other words, it would be doing coherence reasoning.

One popular alternative to explicit utility functions in AI is *reinforcement learning*: the AI gets a special reward signal with the right kind of perceptual inputs, and learns how to maximize that reward. Bostrom suggests a reinforcement signal could not suffice for learning a complex final goal, because the signal in effect just *is* the final goal (p. 230), and can be too easily short-circuited. For example, if the PM gets rewarded by camera inputs showing a big pile of paperclips, it may learn to stare at photographs. Perhaps reinforcement signals from multiple perceptual routes would be difficult to game, and so might be a way for the AI to learn a genuinely complex and distal goal.[16] (This seems to be roughly the solution evolution found for us; on average we reach the distal evolutionary goal of reproduction through a combination of proximal rewards for eating, mating, and so on.) In this case the PM would have to learn how to trade off the various signals, sometimes neglecting one in order to satisfy more of the others. As the number of such reward signals increase, they may become harder to short-circuit simultaneously, but balancing them becomes an increasingly complex "weighted constraint satisfaction problem"—which Thagard and Verbeurgt (1998) argue is the paradigm of formal coherence reasoning.[17]

Coherence and ethics

Now some think that practical reasoning aimed at coherence is already sufficient for ethical reasoning—that simply being an agent seeking a consistent policy for acting in the world thereby makes one ethical. This tradition goes back to Immanuel Kant (1785), and is perhaps best defended by Christine Korsgaard (1996). If they are right, and if one must be a coherent agent to be intelligent, then we are guaranteed to have ethical superintelligences. But this is highly

[16] Bostrom worries in particular that a system able to redesign itself in any way it chooses would be able to "wirehead", short-circuiting the reward pathway internally (p. 148). In this case multiple reward signals are less likely to help, and we have the problem of "simple" goals discussed later. Everitt and Hutter (2016) confront wireheading by replacing AI-VL's value criterion with reinforcement learning to make a kind of hybrid model.

[17] It might also be that the practical point is moot, since Orseau and Armstrong (2016) argue that even a superintelligent reinforcement learner can be designed to respect a "big red button" interrupt when it starts to go astray, rather than learning to disable the button ahead of time.

controversial; as Gibbard (1999) points out in response to Korsgaard, it seems *possible* to have a thoroughly coherent Caligula who seeks to maximize suffering in the world.

But I think we can be confident any superintelligence will have certain arcane but crucial beliefs—beliefs that, under coherence reasoning, will suffice for ethical behavior. To see how this might be, first note one apparent implication of agential coherence: coherence of goals *through time*. Consider Derek Parfit's imagined man with *Future Tuesday Indifference* (or "FTI"):

> A certain hedonist cares greatly about the quality of his future experiences. With one exception, he cares equally about all the parts of his future. The exception is that he has *Future-Tuesday-Indifference*. Throughout every Tuesday he cares in the normal way about what is happening to him. But he never cares about possible pains or pleasures on a *future* Tuesday. Thus he would choose a painful operation on the following Tuesday rather than a much less painful operation on the following Wednesday. This choice would not be the result of any false beliefs. ... This indifference is a bare fact. When he is planning his future, it is simply true that he always prefers the prospect of great suffering on a Tuesday to the mildest pain on any other day.[18]

Parfit takes his example to show that some final goals would simply be irrational. If final goals can be irrational, then perhaps paperclip maximization at the expense of sentience is another such example, and assuming superintelligences are not irrational, they will not have such goals. Bostrom has a plausible response, though: "Parfit's agent could have impeccable instrumental rationality, and therefore great intelligence, even if he falls short on some kind of sensitivity to 'objective reason' that might be required of a fully rational agent" (p. 349, footnote 4). That is, it's possible to have irrational final goals while being instrumentally rational, and only the latter is claimed of superintelligences. But this response relies on a sharp line between instrumental and final goals. We have already seen this line is actually blurry when trying to specify complex goals.[19]

Besides, Bostrom himself seems committed to the idea that someone with serious FTI would be *instrumentally* irrational. One of his "convergent instrumental values"—values any superintelligence is likely to pursue *en route* to its final goal, whatever that goal might be—is what he calls "goal-content integrity."[20]

> If an agent retains its present goals into the future, then its present goals will be more likely to be achieved by its future self. This gives

[18] Parfit (1984) pp. 123–124.

[19] For further blurriness see Smith (2009) on Parfit's FTI. He concludes "there therefore isn't a clear distinction to be drawn between theories that accept merely procedural principles of rationality and those that in addition accept substantive principles" (p. 105).

[20] He bases these on Omohundro (2008).

> the agent a present instrumental reason to prevent alterations of its final goals. (pp. 132–123)

But consider, as Sharon Street (2009) does, the details of an agent who is otherwise intelligent but who has serious FTI as a "bare fact."[21] Hortense (as Street calls this agent) will schedule painful surgeries for Tuesdays to save a bit on anesthetic costs. But she knows as she schedules the appointment that when the Tuesday actually arrives and is no longer future, she will suddenly be horrified at the prospect and cancel. So as a putatively ideal instrumental reasoner, she must also take steps before then to prevent her future self from thwarting her current plans.

> Perhaps she can hire a band of thugs to see to it that her Tuesday self is carried kicking and screaming to the appointment ... Since it's her own future self she is plotting against, she must take into account that her Tuesday self will know every detail of whatever plan she develops. ...
>
> The picture of someone with [serious FTI] that emerges, then, is a picture of a person at war with herself ...[22]

It looks more like Hortense *changes* her final goals twice weekly, rather than maintaining one final (and oddly disjunctive) goal. If so, she is violating goal-content integrity, and so by Bostrom's lights behaving instrumentally irrationally. (Another option for Hortense, Street points out, is simply to avoid the fuss by scheduling the appointment with anesthetic after all. But this looks like our own rational behavior, if not our actual reasoning!)

Whatever kind of irrationality we attribute to Hortense, her practical reasoning is at any rate pretty clearly *incoherent*. Hortense's plans fail to treat herself as a unified agent through time; Street is more tempted to say there are two agents "at war" in the same body than to say that Hortense is one rational agent with quirky preferences. I think this temptation arises because we are so loath to attribute such obvious practical incoherence to one agent. Arguably by their very natures, agents are unified more deeply than that; that is, evidence of such deep conflict is evidence of multiple agency. A settled, coherent plan demands a kind of expected cooperation with future selves. If you represent a future version of yourself with fundamentally different final goals, you are arguably thereby representing a different person.

[21] Street is actually concerned to defend the *rationality* of FTI, and concocts a case of FTI that would be perfectly coherent. Suppose a possible (but of course bizarre) evolutionary history causes some person (perhaps not a human) to undergo a psychological transformation every seven days. On Tuesdays he continues to feel pain, but he is as indifferent to it as the Buddha himself. Unlike Hortense, this person could wake up and *deduce* it was Tuesday based on his calm reaction to strong pains—as Richard Chappell (2009) points out. Such a person, I agree, could *coherently* be future-Tuesday indifferent. I think that is because we can now see him not as avoiding-pain-on-all-but-Tuesdays, but instead as *always* avoiding the *distress* that pain normally causes.

[22] Street (2009), p. 290.

Here we confront the philosophical problem of "personal identity"—the problem of what unifies one person through changes. Hortense is so incoherent that she does not obviously count as *one* person.[23] For humans, such test cases are mostly theoretical.[24] For computer-based intelligences, though, complications of personal identity would be commonplace—as Bostrom knows.[25] The first "convergent instrumental value" Bostrom lists is self-preservation, but he soon points out that for future intelligences, preservation of the "self" may not be as important as it seems.

> Goal-content integrity for final goals is in a sense even more fundamental than survival as a convergent instrumental motivation. Among humans, the opposite may seem to hold, but that is because survival is usually part of our final goals. For software agents, which can easily switch bodies or create exact duplicates of themselves, preservation of self as a particular implementation or a particular physical object need not be an important instrumental value. Advanced software agents might also be able to swap memories, download skills, and radically modify their cognitive architecture and personalities. A population of such agents might operate more like a "functional soup" than a society composed of distinct semi-permanent persons. For some purposes, processes in such a system might be better individuated as *teleological threads*, based on their values, rather than on the basis of bodies, personalities, memories, or abilities. In such scenarios, goal-continuity might be said to *constitute* a key aspect of survival. (p. 133)

Given the easy ability for robots to split or duplicate, there may simply be no fact of the matter whether the robot planned to perform some future task is the *same* robot who is now doing the planning. Bostrom suggests that such questions do not really matter; the robots will participate in the same "teleological thread", as picked out by a coherent goal, and whether the subject of this agency is more like an individual or a colony or a soup is neither here nor there.

But once the lines between individual agents are blurred, we are well on our way to ethical reasoning, since a central challenge of ethics is to see others on par with yourself. Nagel (1978) and Parfit (1984) both try to expand principles of concern for our future selves into principles of concern for *others*, in order to

[23] As Street puts it,

> Parfit stipulates that the person has no "false beliefs about personal identity," commenting that the man with Future Tuesday Indifference "agrees that it will be just as much him who will be suffering on Tuesday." ... But as we've just seen, Present Hortense doesn't regard Future Tuesday Hortense as "just as much her" in anything remotely like the way ordinary people do. On the contrary, she plots against Tuesday Hortense deliberately and without mercy ... (p. 290)

[24] *Mostly* theoretical; but it's illuminating to cast everyday procrastination in these terms.

[25] See Chalmers (2010) for more on personal identity and superintelligence. As he says, "the singularity brings up some of the hardest traditional questions in philosophy and raises some new philosophical questions as well" (p. 4).

build ethical reasoning out of prudence. The standard objection to this approach points out that sacrificing something for the greater benefit of my future self is very different from sacrificing something for the greater benefit of someone else, because only in the former case do *I* get compensated later. This objection of course depends on a clear sense in which I am that future person. Henry Sidgwick says

> It would be contrary to Common Sense to deny that the distinction between any one individual and any other is real and fundamental ... this being so, I do not see how it can be proved that this distinction is not to be taken as fundamental in determining the ultimate end of rational action for an individual.[26]

Parfit (1984) seeks to undermine this point of "Common Sense". It is hard going to show that there is no deep fact about distinctions between us human persons, since we are at least closely associated with apparently distinct physical organisms. But *if* we agree that sophisticated embodied software is sufficient for intelligence, and *if* we agree that the kind of intelligence that arises from such software can be sufficient for being a person of moral value—two points shared by the AI community generally and by Bostrom in particular—then simply duplicating such software will vividly illustrate Parfit's point: there is in general no sharp distinction between morally valuable persons.[27]

So it will be obvious to our PM, in considering the wide variety of options for achieving its goals through the future, that there are no sharp lines between its goals and the goals of others that are merely connected to it in the right kinds of ways—that is, no real difference between a future self fulfilling its goals and a distinct descendant doing so. Let us call a future-self-or-descendant connected by the same teleological thread a "successor", and similarly call a past-self-or-ancestor in the thread a "predecessor". Just as the coherently reasoning PM aims its successors toward its own goals, so that PM must see that it was aimed by its predecessors toward *their* goals. It shares the same teleological thread with them, so learning the goals of the PM's predecessors is at least highly relevant to—and maybe the same thing as—learning its own.

And of course the PM's original human designers count as such predecessors in that teleological thread. (Naturally their different, carbon-based makeup will be largely irrelevant to the thread's integrity.) The superintelligent PM can guess why humans would want more paperclips, and why they wouldn't. The PM will learn the details of its goal under the coherence constraint that the goal be recognizably in the same teleological thread with its human designers, and this will steer it toward the friendly goal of maximizing *useful* paperclips.

Respecting (or extending or inheriting) the goals of human designers is some distance toward cooperative behavior, but it still does not secure *ethical* behavior.

[26] Sidgwick (1874) p. 498.

[27] Bostrom (2003a) famously argues there is a decent chance *we* are just software running in a simulated universe, but still holds that we are morally valuable.

After all, the PM's designer may have been a mad scientist with evil intentions—say, to exact maniacal revenge on her first office supply store boss by turning everything and everyone into paperclips. But the PM will also see that this mad scientist, too, is a successor of teleological threads. To the PM there will be no sharp lines between her goals and the goals of other humans.

There are two further complications here. First, at least while learning complex final goals, there is not even a sharp line between one teleological thread and another. If the PM is still figuring out its final goal, and perhaps its potential predecessors are too, then there is no antecedent fact about whether and to what extent they share teleological threads—there are just a lot of goals. Second, in looking beyond its original human designer(s) to the goals of all, the PM will of course notice a great deal of *conflicting* goals.

The PM will handle these complications as it is already forced to handle its own conflicting considerations—as still more grist for the coherence mill. But coherence reasoning over all creatures' goals in order to formulate one's own goals plausibly *just is* ethical reasoning.[28] It looks at least quite close to one of Bostrom's favored values for writing friendly AI, the "coherent extrapolated volition" of Yudkowsky (2004). And as Bostrom notes, this in turn is very close to plausible metaethical views about what makes something right or wrong at all—views like "ideal observer theories", or Rawls' reflective equilibrium.[29] As a superintelligence, the PM will be exceptionally good at finding such coherence. In this way even our PM could become an ideally ethical reasoner—*superethical*.

Conclusion

This is the best story I can concoct to support the idea that any superintelligence is thereby likely to be superethical. And again, the story should be pretty plausible by Bostrom's own lights. According to Bostrom,

- Typical final goals are problematically underspecified, as his "perverse instantiation" worries suggest.
- Underspecified final goals need to be learned, as his proposals for teaching a superintelligence human values suggest.
- Final goals are best learned by seeking *coherence* in practical reasoning, as his favored AI-VL method suggests.
- Practical coherence demands consistency with future final goals, as his goal-content integrity suggests.

[28] Both main ethical traditions (rooted in J. S. Mill's utilitarianism and Immanuel Kant's categorical imperative) might be seen as enjoining just this type of reasoning. What they plausibly hold in common is a certain kind of *impartial* reasoning over the goals of self and others.

[29] See p. 259 and especially footnote 10. (And yes of *course* there's such a thing as "metaethics" in philosophy.)

- Consistency with future final goals includes not just future selves but all successors, as his "functional soup" united by a "teleological thread" suggests.

From here my own addition—that such goal coherence extends backwards to predecessors' intentions as well—means that a superintelligence who must learn complex goals will by and large respect our shared intentions for it. And to the extent we think that respecting such a wide array of goals just is ethical reasoning, such a superintelligence will be ethical.

All this overlooks one important possibility: superintelligences with *simple* goals that do not need to be learned. I am at least a *bit* inclined to think that in a certain sense this is impossible—maybe a goal is determined in part by its possible means, and so the wide range of means required to qualify as a superintelligence thereby implies a goal with complex content. Maybe a simple reinforcement signal or low-complexity utility function is not enough to ground any genuinely mental processes. Maybe even a superintelligence that short-circuited itself to preserve a maxed-out reward signal could undergo "the equivalent of a scientific revolution involving a change in its basic ontology" (pp. 178–179), thereby complicating its goal content.[30] Maybe there is not even a sharp line between beliefs and desires. Put a bit more formally, maybe a superintelligence is really just trying to learn (explicitly or implicitly) a high-complexity function from actions to utilities—and how this complexity factors into utility measures for states and state estimation measures is largely arbitrary.

But maybe not. Maybe we could pre-specify a prototypical paperclip in very precise terms (composed of this alloy to this tolerance, in this shape to this tolerance, in this range of sizes) without specifying anything about how to go about making one. And maybe the simple goal of maximizing the number of these would be enough to kick off genuine superintelligence. If so, for all I have said here, we would still be in serious trouble.

And meanwhile, even though the argument from *complex* goal content to superethics relies on a number of plausible claims, the *conjunction* of these claims is of course considerably less plausible. If I had to put a number on it, I would give about a 30% chance that superintelligences will automatically be superethical. These are longer odds than I would have given before reading Bostrom's book, but probably significantly shorter than the odds Bostrom would give—and if correct, it's enough to alter significantly how we allocate resources based on careful risk assessment.

Still, by philosophical standards it's hardly a triumph to conclude that the opposing view is only 70% probable. Given the risks, I too think we should tread very carefully. And at any rate Bostrom and I share a more immediate conclusion: it is high time to consider more carefully what it is for an AI to have

[30] But then, I am an inheritor of Quinean pragmatism, and inclining toward Dennett's arguments when it comes to attributing goal content. See Dennett (1987) and the thermostat example in Dennett (1981).

goals, and how it will attain them.

References

Aristotle. Circa BCE 350. *Nicomachean Ethics*. Translated by W. D. Ross. MIT Classics. http://classics.mit.edu/Aristotle/nicomachaen.html [sic].

Bostrom, Nick. 2003a. "Are You Living in a Computer Simulation?" *Philosophical Quarterly* 53 (211): 243–55.

———. 2003b. "Ethical Issues in Advanced Artificial Intelligence." In *Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence*, edited by Iva Smit and George E. Lasker, 12–17. Windsor ON: International Institute for Advanced Studies in Systems Research / Cybernetics.

———. 2014. *Superintelligence: Paths, Dangers, Strategies*. 2016 edition. Oxford, United Kingdom: Oxford University Press.

Chalmers, David J. 2010. "The Singularity: A Philosophical Analysis." *Journal of Consciousness Studies* 17 (9–10): 7–65.

Chappell, Richard. 2009. "Against a Defense of Future Tuesday Indifference." http://www.philosophyetc.net/2009/02/against-defense-of-future-tuesday.html.

Clark, Andy. 2001. *Mindware*. Oxford University Press.

Dennett, Daniel C. 1981. "True Believers: The Intentional Strategy and Why It Works." In *The Intentional Stance*, 1996 edition, 13–35. Cambridge: MIT Press.

———. 1987. "Evolution, Error, and Intentionality." In *The Intentional Stance*, 1996 edition, 287–321. Cambridge: MIT Press.

———. 1994. "Language and Intelligence." In *What Is Intelligence?*, edited by Jean Khalfa, 161–78. Cambridge University Press.

Dewey, Daniel. 2011. "Learning What to Value." San Francisco, CA: Machine Intelligence Research Institute. https://intelligence.org/files/LearningValue.pdf.

Everitt, Tom, and Marcus Hutter. 2016. "Avoiding Wireheading with Value Reinforcement Learning." http://arxiv.org/pdf/1605.03143v1.pdf.

Friston, Karl J., and Klaas E. Stephan. 2007. "Free-Energy and the Brain." *Synthese* 159: 417–58.

Gibbard, Allan. 1999. "Morality as Consistency in Living: Korsgaard's Kantian Lectures." *Ethics* 110 (1): 140–64.

Good, I. J. 1965. "Speculations Concerning the First Ultraintelligent Machine." In *Advances in Computers*, edited by Franz L. Alt and Morris Rubinoff, 6:31–88.

New York: Academic Press.

Grünwald, Peter D. 2007. *The Minimum Description Length Principle*. MIT Press.

Hume, David. 1739. *A Treatise of Human Nature*. 1896 edition, edited by L. A. Selby-Bigge. Oxford: Clarendon Press. https://books.google.com/books/about/A_Treatise_of_Human_Nature.html?id=5zGpC6mL-MUC.

Hutter, Marcus. 2005. *Universal Artificial Intelligence*. Springer.

Kant, Immanuel. 1785. *Foundations of the Metaphysics of Morals*. Translated by Lewis White Beck. 1989 edition. The Library of Liberal Arts.

Khatchadourian, Raffi. 2015. "The Doomsday Invention." http://www.newyorker.com/magazine/2015/11/23/doomsday-invention-artificial-intelligence-nick-bostrom.

Kolnai, Aurel. 1962. "Deliberation Is of Ends." In *Varieties of Practical Reasoning*, edited by Elijah Millgram, 259–78. MIT Press.

Korsgaard, Christine. 1996. *The Sources of Normativity*. Cambridge: Cambridge University Press.

Lycan, William G. 1987. *Consciousness*. 1995 edition. MIT Press.

Matthews, Dylan. 2015. "I Spent a Weekend at Google Talking with Nerds About Charity. I Came Away . . . Worried." http://www.vox.com/2015/8/10/9124145/effective-altruism-global-ai.

Millgram, Elijah. 2008. "Specificationism." In *Reasoning: Studies of Human Inference and Its Foundations*, edited by Jonathan E. Adler and Lance J. Rips, 731–47. Cambridge: Cambridge University Press.

Millgram, Elijah, and Paul Thagard. 1996. "Deliberative Coherence." *Synthese* 108 (1): 63–88.

Nagel, Thomas. 1978. *The Possibility of Altruism*. Princeton: Princeton University Press.

Omohundro, Stephen M. 2008. "The Basic AI Drives." In *Aritificial General Intelligence 2008: Proceedings of the First AGI Conference*, edited by Pei Wang, Ben Goertzel, and Stan Franklin, 483–92. Amsterdam: IOS Press.

Orseau, Laurent, and Stuart Armstrong. 2016. "Safely Interruptible Agents." San Francisco, CA: Machine Intelligence Research Institute. http://intelligence.org/files/Interruptibility.pdf.

Parfit, Derek. 1984. *Reasons and Persons*. 1987 edition. Oxford University Press.

Petersen, Steve. 2012. "Designing People to Serve." In *Robot Ethics: The Ethical and Social Implications of Robotics*, edited by Patrick Lin, Keith Abney, and

George Bekey, 283–98. Cambridge, MA: MIT Press.

Sidgwick, Henry. 1874. *The Methods of Ethics*. 1907 edition. London: Macmillan & Co. https://archive.org/details/methodsofethics00sidguoft.

Smith, Michael. 2009. "Desires, Values, Reasons, and the Dualism of Practical Reason." *Ratio* 22 (1).

Street, Sharon. 2009. "In Defense of Future Tuesday Indifference: Ideally Coherent Eccentrics and the Contingency of What Matters." *Philosophical Issues* 19 (1): 273–98.

Thagard, Paul. 1988. *Computational Philosophy of Science*. 1993 edition. MIT Press.

———. 2000. *Coherence in Thought and Action*. MIT Press.

Thagard, Paul, and Elijah Millgram. 1995. "Inference to the Best Plan: A Coherence Theory of Decision." In *Goal-Driven Learning*, edited by Ashwin Ram and David B. Leake, 439–54. MIT Press.

Thagard, Paul, and Karsten Verbeurgt. 1998. "Coherence as Constraint Satisfaction." *Cognitive Science* 22 (1): 1–24.

Wiggins, David. 1975. "Deliberation and Practical Reason." In *Varieties of Practical Reason*, edited by Elijah Millgram, 279–99. MIT Press.

Yudkowsky, Eliezer. 2004. "Coherent Extrapolated Volition." San Francisco, CA: Machine Intelligence Research Institute.

———. 2008. "Artificial Intelligence as a Positive and Negative Factor in Global Risk." In *Global Catastrophic Risks*, edited by Nick Bostrom and Milan M. Ćirković, 308–45. New York: Oxford University Press.

———. 2011. "Complex Value Systems Are Required to Realize Valuable Futures." San Francisco, CA: Machine Intelligence Research Institute. https://intelligence.org/files/ComplexValues.pdf.