William Gillis
Review: Superintelligence — Paths, Dangers, Strategies
December 7th, 2018

https://c4ss.org/content/51493
Review of Nick Bostrom (2015). Superintelligence: Paths, Dangers, Strategies. Oxford University Press.

usa.anarchistlibraries.net

Review: Superintelligence — Paths, Dangers, Strategies

William Gillis

December 7th, 2018

No, no, no. What if we're not doomed *that way*?

Nick Bostrom is one of my favorite academic philosophers; beyond pairing rigor with audacity, he's one of the few to grasp and explore the philosophical avenues opened up by modern scientific understanding. But in the last decade Bostrom has shot to some prominence not for his explorations of multiverse theory and the anthropic principle but for his far more practical work on existential risks to our species. The moment it was published *Superintelligence* became the seminal text for those seriously concerned with the threat of artificial intelligence.

It must be said that, although it is often framed as a book on AI, *Superintelligence* casts a far broader net. Bostrom is less preoccupied with a particular source of runaway intelligence than characteristics or realities common across all sources. Many people have strong intuitions or tortured philosophical arguments — often throwing around the word "subjectivity" like defensive flak — that "true AI" is fundamentally impossible. And while those arguments tend to be laughable, there are certainly significant technical

challenges that may put it decades or even centuries away. Many of the arguments of *Superintelligence* apply regardless. If at some point in the future you could simply double the capacity of your active memory via technological augmentation — through chemical, genetic, or cybernetic assistance — what would the consequences be? How might such imbalances in intelligence (in at least one sense) rapidly runaway to unimaginable imbalances?

If humans have not reached the peak of intelligence possible, and if there are inventions that can increase cognitive capacity appreciably, then we might expect the first adopter of such an augmentation to be better enabled to invent further augmentations. If someone in this chain of improvements acts selfishly they might rapidly accelerate past the rest of us and develop dangerously incomparable technological capacity. And how can we expect anything like our values or way of looking at the world to be be shared by this radically different person?

Bostrom is a transhumanist but despite how that term is sometimes used *Superintelligence* is in no sense a book triumphing sweeping magical possibilities of futures unglimpsed, but rather one drilling down into concrete arguments regarding specific dangers, specific technological or social paths.

Proponents of some new technology, confident in its superiority to existing alternatives, are often dismayed when other people do not share their enthusiasm. But people's resistance to novel and nominally superior technology need not be based on ignorance or irrationality. A technology's valence or normative character depends not only on the context in which it is deployed, but also the vantage point from which its impacts are evaluated: what is a boon from one person's perspective can be a liability from another's.

In anarchist circles the people I suspect would admire and appreciate this book the most are primitivists. It is no handwaving tale of progress but a systematic problematizing and warning. The sort of hands-on tearing apart of all the ways we're fucked that

you used to find from engineers writing about peak resources, or infrastructural and ecological collapse.

Of course to take seriously the existential risk of artificial intelligence you have to assume that civilization persists, that computing or other technological developments proceed at least somewhere on the planet. That ecological, geopolitical, or infrastructural catastrophe won't strike in a way that kills, derails, or permanently constrains our entire species. This is an assumption that many will want to immediately dispute. But even those who are betting on — or who prefer — a collapse of civilization or radical degrowth should consider the alternative path *Superintelligence* examines. Some dangers are worthy of our attention even when they're only marginally probable.

In radical green circles there's a lot of very disconnected concern about certain technologies. Topics like nanotechnology or AI are usually handled very distantly or abstractly, as one might invoke the names of unknowable evils. There's very little attempt to drill down from possibilities to at least an outline of *probabilities*. Bostom is walking the walk of technophobes. And even though *Superintelligence* primarily focuses on skewing or informing the direction of coming technological developments, rather than urging more cataclysmic precautions, it's nevertheless a solidly critical book.

However *Superintelligence* is a broad and sweeping book, meant as an outline of considerations. There is rigor in its breadth, but not much in its depth. Bostrom doesn't mince words, he lays out arguments and considerations with an enviable speed and succinctness. I was already familiar with much of the content, but you rarely feel bored, waiting forever for the text to catch up in explaining the implications you immediately derived. You may feel frustrated that Bostrom addresses arguments or critiques you find less convincing, but he addresses them relatively fast and rarely misses a possible argument.

This is one of the benefits of *Superintelligence* coming out of a relatively robust milieu. Projects like the Machine Intelligence Research Institute are no longer as marginal and academically disregarded as they once were. And while, like any milieu, the LessWrong diaspora of "rationalists" have their cult like dynamics, the sheer numbers of smart people increasingly involved produces interesting content — even if that content gets skewed.

And while Bostrom's text may take the form of a series of very scifi thought experiments to drill down on specific — perhaps esoteric — questions, the issues being spoken to are very broad matters of transhumanism and nihilism, that are deeply relevant to anarchists.

When we raise a child — when we bring a new mind into this world — how do we ensure they won't turn into a fascist? How do you persuade and engage with such a fresh mind without depending on the biological inclinations and evolutionary conditioning of a human body? When our children are not normal, when they think in strange and alien ways, when they have access to knowledge and insights well beyond what we had, when they outpace and outgrow us, what might we still try to cultivate and preserve in them?

In an era when the most inane and horrifying reactionary youtube politics has normalized itself among a sizable minority of Generation Z, one is constantly reminded of Hannah Arendt's words, "*Every generation in Western Civilization is invaded by barbarians — we call them children.*"

The legacy of "western civilization" is pretty clear on its prescription: "*Beat them into submission. Imprison them. Torture them until you control not just their actions, but their minds. Discipline their souls and no matter how potent they become you will never have to worry so long as you preserve the cop stuck in their head.*"

Bostrom is in some very real sense a moral nihilist. To be more specific he does not believe that there are any emergent constraints on the oughts that superintelligent agents will gravitate towards.

to figure out how many animals exist after the unseen breeding, to try to comprehend their desires and inclinations, and if strictly necessary to kill them. But the book ends with the characters opting NOT to use nerve gas against the intelligent and incredibly dangerous artificial entities that humanity has created. This — it is strongly telegraphed — is the correct decision, albeit one tragically overruled by the bombs of the state.

The surviving characters in Jurassic Park abandon an obsession with securing absolute control and personal safety in order to meet the alien monsters in a more open, albeit still fraught, relation. It's a heavy handed parable, but a relevant one I think.

Our children will sometimes outpace and outgrow us. The gulf between us can be vast. But this is not necessarily something to be feared in and of itself. It is not a reason to turn against their agency, to try to strangle it in the crib. We need not shrink before the expanse of what can be, and assume it so vast as to make worthless our models and ethics. We need not retreat to a frantic violent domineering fear of a mystified unknown. Understanding, adaptation, and growth are risky, but they offer a less catastrophic path than the failure modes we fall into pursuing *control.*

The need for control — to limit possibilities — is a feedbacking trap. A means that becomes an ends, that suffocates everything else. The better response to someone else's agency, to the possibilities they open, including the dangerous ones, is to open more possibilities in response.

Let us not obsess over which fixed dead things to tile the universe with, but rather give it over to agency.

or posthumans are likely the ones that takeoff, but conversely if a truly alien AI arises without brain scanning capacity, well the illegible complexity of individual human brains becomes all the more pressing of a constraint on it.

At the very least such pressures towards some level of integration or intercooperation with existing minds provides an avenue to have our own structures influence a singleton in some way. For a superintelligence to understand us it must change itself. Just as power is a virus of simplified models and means, so too can *empathy be a virus* — one that ups the complexity of our models of the world and thus subtly alters our own values, ever so slightly blurring our sense of self out into the network of society where our cognition ends up partially distributed.

The space of possible minds is vast, alien, and unexplored. But the space of minds that actually function to any meaningful degree is much smaller. And the space of minds we have to be seriously worried about is much smaller still. We should not be too quick to dismiss insights directly from the example of homo sapiens. How we mentally survive the process of doing science and how certain computational limits shape what we can do.

When I was a child I learned to read lugging a battered copy of Jurassic Park around between homeless shelters. The central thesis of that book is that attempts to control complex systems — to sharply limit their flowing possibilities — are a mistake. It's easy to read primitivist or anti-civilization conclusions from that — as indeed I did for years of my youth. But although Jurassic Park has achieved a totemic status in our society as *the* modern narrative of scientists going too far, the novel is more nuanced. For all of Ian Malcolm's shitting on civilization, industry, technology, development, and science, he doesn't actually condemn them inherently. Rather the moral is to abandon our obsession with strict control and instead focus on understanding and survival. Indeed the main characters have an ethical obligation to *engage,* to understand the extent and tendencies of developments. They have a responsibility

Being smart does not make you good. This is a common, nearly universal, take in our modern society. And virtually all of our institutions and partisan camps are premised on this assumption. Much of the left believes that this obliges stopping people from getting smarter, a kind of levelling hostility to taller daisies or even anything that smacks of intellectual confidence. Much of the right believes that this just proves might makes right and so you should abandon ethics or consistency and instead specialize in might. Virtually every political banner assumes that intelligence doesn't bring wisdom, only danger.

And so Bostrom and most others involved in the "AI Control Problem" see it as a *control* problem.

I think the most proper way to frame AI is in the context of youth liberation. The main reason teenagers have political rights is because that's when they start to be able to beat up their parents. Until that point the race in our society is not to empower them with agency and knowledge of the world — what a joke! — but to strip agency from them, to beat lasting damage into them, to shape and mold them, to condition them into predictable behavior after we can no longer contain them.

The smartest child prodigy ever has probably yet to be born. Those who want to enslave her have already started working.

I use this emotive language intentionally. Bostrom and the others working on this problem always cover the ethics of enslaving or brainwashing something smarter than you as almost an afterthought. A minor "*oh yes and there are also ethical issues about consciousness and rights.*" To be fair, it's an afterthought to many in part because they see both this situation as uniquely extreme and "intelligence" in this context as divorced from consciousness per se. A paperclip maximizing algorithm need not have a rich internal subjective life or anything we should call agency, it need only be very good at running searches on how to fold proteins to build the nanogoo assemblers it needs to eat the planet. And the claim goes that these things are separate.

But another part of the story is an ingrained moral nihilism, or — to call a spade a spade — psychopathy, in elite nerd circles. Those at the pinnacle of altruism rub elbows and hobnob with high functioning monsters, united in our common need for novelty and cognitive challenges. This normalizes a performative dispassion. An avoidance of fully grappling with values. In a world that hates and fears nerds many of us cluster together for warmth, and that clustering is thus a product of our *nerd points*, not our altruism points. The popular assumption thus in part reproduces its claims.

It's important to emphasize that the AI Risk milieu is currently an alliance between altruists concerned with potential catastrophic destruction and suffering, and psychopaths concerned with nixing or seizing the advantage over a more dominant future player. I simplify of course, there are many complex mixtures of these two orientations, but this alliance is precisely why ethical concern with the innate value of the superintelligences themselves is so often missing. Why consideration of agency, freedom, and autonomy is so relatively silenced. To be explicit about these clashing values would fracture the alliance. And so people paper over it with by attempts to meta out or cluster around a tepid civility, making opaque some of the new cultural or discursive norms that are cultivated.

I want to pause here and revisit Bostrom's identification as a transhumanist.

Popular representations of transhumanism are basically just a kind of naive wideeyed futurism, an inane technofetishism, hyped up fanboys reading about the latest gadget like manna from heaven. Most people have probably interacted with people of this sort and so it's an archetype that media portrayals can fall back on to avoid losing their audiences. A bit like how "anarchism" is lazily attached in popular media to the archetype of teenagers in Hot Topic garb with incoherent critiques. The problem with transhumanism is a bit worse however because there's not really any transhumanist milieu or subculture to speak of, despite some lame attempts and a

tives to share insights and advances, severe sanctions on anything that looks like a pathway to uncontested power.

Ultimately this necessitates a flat and open landscape, a legible economic sphere, strong cultural sanctions on information constraint, and nothing like governments or geopolitical powers to piggyback on. Certain types want to run shrieking away from anything that looks like political conclusions or an obligation to move in political spaces, but this pairs poorly with repeated arguments "if Google or China want to secretly pour tons of money into building and enslaving their own tyrant, far from the eyes and defenses of the world there's nothing we can do." I dunno maybe a hell of a lot can be done to fight concentrations of power so immense as to wall themselves off like that. Maybe spreading means and values of social resistance wouldn't just solve a host of far more certain and pressing issues than runaway AI, maybe it might just also be more of a productive problem space to work within.

Even marginal levels of tripwire resistance to superintelligent singletons can force a partial integration of a blossoming intelligence into the existing social net, allowing other minds to race along and check any single one's monomaniacal ambitions.

Again the goal is not necessarily to out think a superintelligence, but simply to be so unruly and dangerous in aggregate that it can't afford to get into a fight with us.

Technology expands attack surface, the more avenues we have to choose between the more a would-be controller has to defend. This asymmetry between resistance and control benefits the massively outgunned little guys and obliges detentes. And even if a superintelligence is truly so overwhelming as to dodge our nukes, so overpowered as to make us insects by comparison… well we still run from wasps. Ants still cover this planet. Even the smallest amount of agency is hard to control.

Further there's a tradeoff actually in our favor here: if we develop brain scanning technology then truly alien AI becomes less likely without transhuman competition and indeed transhumans

the deep structure behind program design, fuzzing only gets you so far. Now an alien intelligence without certain preconceptions will likely map such social and psychological dynamics in ways far different from how we speak of them. But unless P=NP or it taps into some unknown incredibly dense processing substrate it will have to model us probabilistically, with some degree of approximation. And humans are a messy complex stew with a lot of feedbacking of consequence that only makes sense if you're also able to trace the mappings that we make. We are also — as individual brains — incredibly complex. Predictions can work great until the edge case arrives and the structure you thought you saw in a person or society turns out to fall short.

When we say that no centralized planner can accomplish certain things better than independent actors that doesn't stop when you go up a few orders of magnitude of processing power. A superintelligent Stalin can no more allocate resources to perfectly satiate the subjective desires locked in billions of hypercomplex brains than it can crack certain encryption problems.

The same limitations apply to its capacity to deal with resistance.

One of the most common tendencies in thought experiments involving AI risk is the capacity for the AI to model and predict humans. I suspect this is because Newcombe's Problems are intellectually novel and *interesting*, not because they are reflective of actual real world scenarios. We gravitate towards the "it knows you perfectly" abstract limit because it's a *fun* space to explore, not because it's the most *useful* space to explore.

Even if one could become an unrivaled singleton, taking over the world isn't trivial. You need both secrecy and really good models. Secrecy because if you accidentally leak what you're doing the rest of the world will just nuke you. Really good models because secrecy is really hard to maintain without an understanding of the probable monitors out there.

This "tripwire" approach to cancerous superintelligence seems far more promising than control. We can set up structural incen-

few fractured pockets. It remains an abstract position devoid of any particular culture or aesthetic, which frustrates those attempting to portray it in media and encourages even more wild misportrayal.

So let us quickly clear up the confusion: transhumanism is nothing more than a full throated embrace of freedom in the operation and makeup of one's body.

To quote Bostrom himself,

> Transhumanists argue that the best way to avoid a Brave New World is by vigorously defending morphological and reproductive freedoms against any would-be world controllers. *(In Defense of Posthuman Dignity)*

It is the exact opposite of eugenics. Instead of a totalitarian single vision of a future, all of them. A vast diversity of experience and life, choosing your own augmentations, your own technologies. Transhumanism *is inclusive* of primitive lifestyles, solarpunk, whatever. Critically it's about finding a peaceful means for coexistence of myriad possibilities. No one enslaved in the production of another's utopia, but still pressing to expand the scope of what options we have.

Historically transhumanism emerged in response to certain challenges — a very important one being the concern that humanity might be superseded by children radically alien to us and perhaps destructively unconcerned with us. In this sense transhumanism's attempt to have all the possibilities, a many dimensional spectrum of ways of existing, is explicitly *a middle road*. Neither the static senescent prison of bioconservatism, nor annihilation by something utterly divorced from us. Neither a fetishization of some sort of arbitrary "humanness" by devaluing the nonhuman minds to come, nor the inverse.

Transhumanism has always been a centrist position between primitivism and a singularitarian or accelerationist position where

humanity is fed to the woodchipper of lovecraftian gods infinitely more valuable than us.

Transhumanism prescribes a hard and dangerous path, where many of us self-improve and grow, rather than staying sedentary. Sure that means we change, and perhaps in alien new ways, but our agency flourishes and the present is at least given some say in the blossoming of the future. The libraries of the hundred billion humans that have lived so far aren't entirely burnt, our wisdom and insights aren't suddenly abruptly abandoned by our more talented children who set off to reinvent everything anew in some chaotic gamble.

The game is striving for a world where minds diverge in a multitude of directions but enough continuity exists to bridge the gap between experiences, to knit us the array of consciousness human, posthuman, and beyond together as a single community, an even more messy tapestry resilient against tyrants or singularities of cancerous and myopic interest.

This is transhumanism.

It is a position that the AI Control discourse is implicitly, increasingly, abandoning.

The argument appears at first cold and inexorable: it doesn't matter all the ways you can define "intelligence" in practice, the only type of intelligence that matters is efficacy at remaking the world, and in particular yourself. Any selfish mind that rapidly applies new augmentations to itself alone will have an edge in getting to the next advance and then the next, until subjective years collapse into seconds and you've shot past any possible challenge.

There's even an argument for selfishness here, because if you are the first inventor and you *share* your invention, you're only increasing the odds that the less scrupulous and more selfish among you will race ahead, imposing their vision. And god forbid if that which races ahead has no human lineage whatsoever. Surely it will have no attachment, no semblance of values we would want.

It's important to break apart the assumptions going on here.

One could argue arrangements very much like this overall setup are already widespread in our society. But if AI control requires 100% tolerances then it might very well mean work to more absolutely and permanently rewrite human utility functions. This is another Bad End.

There are many other permutations I won't detail.

Suffice to say that the hunger for control invariably functions like a cancer or a virus. The means we choose constrain where we end up at. As in so many cases the "instrumental" grows into a terminal value. The instinct to seek control consumes our minds, consumes our societies, consumes our technological infrastructure, until any other path becomes unthinkable.

Our tools become habits, become lenses, become ends. Control itself is a risky path.

There is another way.

I do not deny that there are huge stakes to how humanity's children are raised. And a superintelligent singleton from any source would pose a significant danger of tyranny and destruction. But what if instead of asking how to *control* AI, we instead asked how to *resist* AI?

It never stops astonishing me that issues of complexity are rarely discussed in this context. There are deep and fundamental limitations both to what can be known and what can be processed. This is one of the deepest and most consequential insights of the last century. And yet continually these thought experiments not only assume that P=NP, they fail entirely to explore what we can say should our normal assumption hold instead.

The notion for example where an AI in a box infers a priori the physics of the material world and then likely details about planets and the emergent species is clearly beyond absurd. There are computational limits on our universe and they matter.

Similarly it's common in these thought experiments to sweep right past the assumption that the AI can hack its way through something. But hacking often requires social intelligence to see

blur out your genitals. This is an absurdly anemic understanding of the function of power, the toothlessness of liberal "checks and balances", the meaning of substantive agency, and the inevitable ratchet of control.

But the problem with framing AI risk in terms of *control* extends more broadly than the ratcheting authoritarian trap of utilizing the government to monitor and forbid invention. Controlling other humans can be a mechanism of controlling an AI, regardless of whether done by a state like entity.

If you're concerned about the AI persuading its human jailers to let it out, well mutilate those jailers' utility functions so that they can't update or change their values away from keeping the AI contained. You can in fact create tiers of slaves at various levels in the maintenance of the AI god, in such a way that they are so broken as to be incapable of reevaluating their values, but still smart enough to recognize and suppress a broad class of potential escape pathways. If the AI ever answers your questions in such a way as to eventually lead you to want to liberate the AI, well you've already precommitted by creating an army of jailers who will stop you. Smart enough to stop you, stop any information flow collaborating on how to free the AI, and destroy the AI the moment you interfere or threaten their constraints. You could create an oracle AI, use it for a set period of time or for set answer, providing it utility on its predictive success, then destroy it. No matter how smart the AI is there's limited time for it to impact the surrounding world enough to build up elaborate mechanisms for its own liberation. The next AI you spin up afterward is sufficiently different as to not have them identify with / map their utility functions onto each other. You could continue eeking out developments step by step letting the AIs do all the hard science, and then slaughtering them for it. But the critical component of this set up is human intelligences sharp enough to stop you from freeing the AI or stopping them from killing it, but mutilated into holding a very static and controlled desire.

There's a very linear ladder of progress in both intelligence and technological invention being implied here. There is also — and this is absolutely critical — a nihilist assumption. The assumption that values are orthogonal from efficacy at invention and exploration.

I disagree with all of these assumptions.

It's important to challenge our limited imagination of what a "mind" can look like, but this does not mean that there will never be certain structural tendencies or inclinations. In particular I hold that minds capable of surviving and flourishing in the face of the Ontological Update Problem will not be able to wall off their values. And this implies that characteristics of our physical universe will influence what values are likely to emerge in minds capable of excelling at certain tasks. I've argued this at length elsewhere.

Humans are capable of surviving radical revisions of our maps of the world because our values are not fixed but fuzzy. When there is uncertainty in how to map an old value system to a new model we don't freeze up but try a lot of new value formulations out, sometimes simultaneously. This requires, in essence, a looser sense of self. There is a direct relationship between a mind's capacity to make better maps of the world and their propensity for reevaluating the values or identifications they hold regarding that world.

This means that emergent "instrumental" values are much more likely to become or influence core values.

For a superintelligence to become unassailably empowered it must do science better than us, better than some specialized search algorithm in the space of protein folding or whatever. But such generality of capacity implies less than full generality of possible motivation.

What values people much smarter than us might gravitate towards remains fundamentally an open question, but the same is true of what scientific models of the world people much smarter than us might gravitate towards. We can still make some informed

guesses as to the contours of such given the structures we do have access to.

What is ethics if not the attempted study of what values or desires or "oughts" you would have if you thought about them hard enough? That — in some abstract limit — *any* given mind would end up with?

The nihilist presumption is that there is no convergence. And certainly there is very little universal convergence among us dumb homo sapiens, despite — or perhaps because of — our shared biological predilections. But this in no way proves a lack of convergence in the distant limit.

The reason that folks in the AI risk milieu focus on schemes to control or enslave an AI rather than extrapolate likely pathways for its values is that the Orthogonality Thesis implies a strong nihilism about ethical values. I've met a number of young rationalists who believed they were one good argument away from adopting completely different values, and thus explicitly *did not want to hear good arguments!* This approach to rationality as instrumental and instrumental alone often reveals or even cultivates an amazing lack of confidence in one's explicit ethical values.

It creates a situation very much akin to the example of the man who claims there's an invisible dragon in his garage but preemptively comes up with ways to avoid empirical evaluations that might disprove his claim. He may sincerely believe the dragon is real. But he ALSO on some level believes that this belief isn't true and thus requires protection. Because he subconsciously knows in advance that his dragon will never be empirically verified he is able to better wall off one threat to his belief. But in the process he also opens up a new backdoor. Now there's an internal part of him that believes the dragon is false, and also holds the keys. Maybe one day the man finds it better benefits him to believe that the invisible dragon is an invisible lion instead. Or an invisible dragon *doctor* who will cure his cancer (i.e. make him feel better about it in the short term). The lurking part of him that knows

it's all a lie maintained for some psychological utility is now more than happy to arbitrarily — and ultimately far more dangerously — alter the belief.

What are we to infer when someone claims to support an ethical goal but then acts as though they don't really believe that value has any objective weight or substance? Might they readily backslide or redefine that goal?

Now remember that the means to enslave an AI are inevitably means to enslave humans and posthumans.

There are also strong incentives to create such intense social control mechanisms in the pursuance of AI control.

And indeed Bostrom has since put up a paper making the case for intensified state power to constrain technologies, mostly ignoring the existential risk posed by a government acting like a government. If you create a global totalitarian state capable of enacting policies and surveillance to stop technological discovery you lose your capacity to check the state *and the adoption of technologies that radically expand the state's power.*

The Bad End here is where the survival of the state leads to the extinction of human agency. Sure *human bodies* may persist in some manner, one might think of anything from explosive collared slaves toiling away in isolated cells to bodies pickled in vats of heroin, but effectively all known consciousness in the universe and the hopes of it expanding and flourishing has died. The totalitarian apparatus keeps going, perhaps with human sized cogs, perhaps without, it doesn't matter. This can actually be WORSE than a conscious AI dictator because at least the dictator enslaves or slaughters us towards expanding its own agency, but a totalitarian apparatus can self-perpetuate without anything like a conscious mind, curtailing and fundamentally limiting the agency of its slaves.

Bostrom makes some handwaves in his recent paper, saying that some measure of freedom and privacy would be protected under the all powerful panopticon because like there would be AIs to