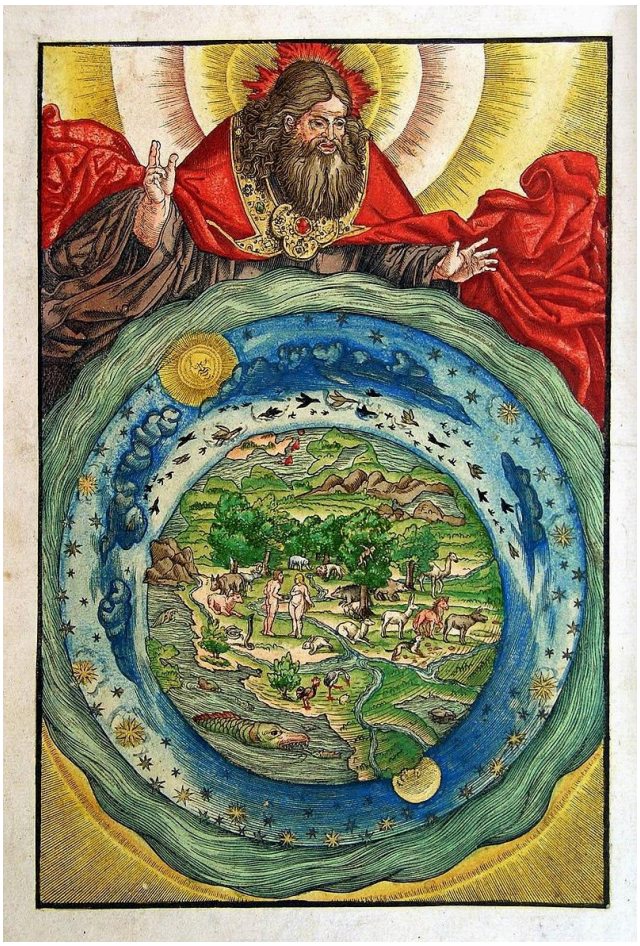
Otherness and control in the age of AGI



Joe Carlsmith

June 18, 2024

Audio version | Web version

Dedicated to Walter Kaufmann

Thanks to Katja Grace, Rebecca Kagan, Will MacAskill, Ketan Ramakrishnan, Anna Salamon, Carl Shulman, and many others over the years for conversation about these topics; thanks to Carl Shulman for written comments; and thanks to Sara Fish for formatting help. I am speaking only for myself and not for my employer.

joecarlsmith.com

Cover painting: "The Creation" by Lucas Cranach (source here).

"With malice towards none; with charity towards all; with firmness in the right, as God gives us to see the right..."

–Abraham Lincoln

Contents

Introduction

I've written a series of essays that I'm calling "Otherness and control in the age of AGI." The series examines a set of interconnected questions about how agents with different values should relate to one another, and about the ethics of seeking and sharing power. They're old questions—but I think that we will have to grapple with them in new ways as increasingly powerful AI systems come online. And I think they're core to the discourse about existential risk from misaligned AI (hereafter, "AI risk").[1]

The series covers a lot of ground, but I'm hoping the individual essays can be read fairly well on their own. Here's a summary of the series as a whole.

- The first essay, "*Gentleness and the artificial Other*," discusses the possibility of "gentleness" towards various non-human Others—for example, animals, aliens, and AI systems. And it also highlights the possibility of "getting eaten," in the way that Timothy Treadwell gets eaten by a bear in Werner Herzog's *Grizzly Man*: that is, eaten in the midst of an attempt at gentleness.
- The second essay, "*"Deep atheism and AI risk*," discusses what I call "deep atheism"—a fundamental mistrust both towards Nature, and towards "bare intelligence." I take Eliezer Yudkowsky as a paradigmatic deep atheist, and I highlight the connection between his deep atheism and his concern about misaligned AI. I also connect deep atheism to the duality of "*yang*" (active, controlling) vs "*yin*" (receptive, letting-go). A lot of my concern, in the series, is about ways in which certain strands of the AI risk discourse can propel themselves, philosophically, towards ever-greater *yang*.
- The third essay, "*When 'yang" goes wrong*," expands on this concern. In particular: it discusses the sense in which deep atheism can prompt an aspiration to exert extreme levels of control over the universe; it highlights the sense in which both humans *and* AIs, on Yudkowsky's narrative, are animated by this sort of aspiration; and it discusses some ways in which our civilization has built up wariness around control-seeking of this kind. I think we should be taking this sort of wariness quite seriously.
- Pursuant to this goal, the fourth essay, "*Does AI risk 'other' the AIs?*," examines Robin Hanson's critique of the AI risk discourse—and in particular, his accusation that this discourse "others" the AIs, and seeks too much control over the values that steer the future. I find some aspects of Hanson's critique uncompelling and implausible, but I do think he's pointing at a real discomfort.
- The fifth essay, "*An even deeper atheism*," argues that this discomfort should deepen yet further when we bring some other Yudkowskian philosophical vibes into view—in particular, vibes related to the "*fragility of value*," "*extremal Goodhart*," and "*the tails come apart*." These vibes, I suggest, create a certain momentum towards deeming more and more agents—including: human agents—"misaligned" in the sense of: not-to-be-trusted to optimize the universe very intensely according to their values-on-reflection.

[1] There are lots of other risks from AI, too; but I want to focus on existential risk from misalignment, here, and I want the short phrase "AI risk" for the thing I'm going to be referring to repeatedly.

And even if we do not follow this momentum, I think it can remind us of the sense in which AI risk is substantially (though, not entirely) a generalization and intensification of the sort of "balance of power between agents with different values" problem we already deal with in the purely human world—a problem about which our existing ethical and political traditions already offer lots of guidance.

- The sixth essay, "*Being nicer than Clippy*," tries to draw on this guidance. In particular, it tries to point at the distinction between a paradigmatically "paperclip-y" way of being, and some broad and hazily-defined set of alternatives that I group under the label "niceness/liberalism/ boundaries."[2] Too often, I think, a simplistic interpretation of the alignment discourse imagines that humans and AIs-with-different-values are both paperclippy at heart—except, only, with a different favored sort of "stuff." I think this picture neglects core aspects of human ethics that are, themselves, about navigating precisely the sorts of differences-in-values that the possibility of misaligned AI forces us to grapple with. I think that attention to this part of human ethics can help us be better than the paperclippers we fear—not just in what we do with spare resources, but in how we relate to the distribution of power amongst a plurality of value systems more broadly. And I think it may have practical benefits as well, in navigating possible conflicts both between different humans, and between humans and AIs. That said, I don't think that "niceness/liberalism/boundaries" is enough, on its own, to ensure a good future, or to allay all concern about trying to control that future over-much.

- The seventh essay, "*On the abolition of man*," examines another version of that concern: namely, C.S. Lewis's argument (in his book *The Abolition of Man*) that attempts by moral anti-realists to influence the values of future people must necessarily be "tyrannical." I mostly disagree with Lewis—and in particular, I think he makes a number of fairly basic philosophical mistakes related to e.g. compatibilism about freedom, to the difference between creating-Bob-instead-of-Alice vs. brainwashing-Alice-to-make-her-like-Bob, and to the sense in which moral anti-realists can retain their grip on morality. But I do think his discussion points at some difficult questions about the ethics of influencing the values of others, including AIs—questions the essay takes an initial stab at grappling with.

- The eight essay, "*On green*," examines a philosophical vibe that I (following others) call "green," and which I think contrasts in interesting ways with "deep atheism." Green is one of the five colors on the Magic the Gathering Color Wheel, which I've found (despite not playing Magic myself) an interesting way of classifying the sort of energies that tend to animate people.[3] The colors, and their corresponding shticks-according-to-Joe, are:

  - *White*: Morality.
  - *Blue*: Knowledge.
  - *Black*: Power.
  - *Red*: Passion.
  - *Green*: . . .

[2]Of course, there are lots of other options as well.

[3]My relationship to the MtG Color Wheel is mostly via somewhat-reinterpreting Duncan Sabien's presentation here, who credits Mark Rosewater for a lot of his understanding. My characterization won't necessarily resonate with people who actually play Magic.

I haven't found a single word that I think captures green, but associations include: environmentalism, tradition, spirituality, hippies, stereotypes of Native Americans, Yoda, humility, wholesomeness, health, and *yin*. The essay tries to bring the vibe that underlies these associations into clearer view, and to point at some ways that attempts by *other colors* to reconstruct green can miss parts of it. In particular, I focus on the way green cares about *respect*, in a sense that goes beyond "not trampling on the rights/interests of moral patients" (what I call "green-according-to-white"); and on the way green takes *joy* in (certain kinds of) *yin*, in a sense that contrasts with merely "accepting things you're too weak to change" (what I call "green-according-to-black").

- The ninth essay, "*On attunement*," continues the project of the previous essay, but with a focus on what I call "green-according-to-blue," on which green is centrally about making sure that we act with enough *knowledge.* I think there's something to this, but I also suggest that green cares especially about "attunement"—a kind of meaning-laden receptivity to the world—as opposed to more paradigmatically blue-like types of knowledge. What's more, I think that attunement is core to certain kinds of *ethical* epistemology, including my own; and it plays a key role in my own vision, at least, of a "wise" future. And while attunement may, ultimately, be made out of red and blue, I think we should take it seriously on its own terms.

- The tenth essay, "*Loving a world you don't trust*," closes the series with an effort to make sure I've given both *yang* and "deep atheism" their due, and been clear about my overall take. To this end, the first part of the essay praises certain types of *yang* directly, in an effort to avoid over-correction towards *yin*. The second part praises something quite nearby to deep atheism that I care about a lot—something I call "humanism." And the third part tries to clarify the depth of atheism I ultimately endorse. In particular, I distinguish between *trust* in the Real, and various other attitudes towards it—attitudes like love, reverence, loyalty, and forgiveness. And I talk about ways these latter attitudes can still look the world's horrors in the eye.

I'll also note three caveats about the series as a whole. First: while I think it likely that ours is the age of AGI—still, maybe not. Maybe I won't live to see the age that I wrote this series for. But I think that much of the content will be of interest regardless of your views on AGI timelines.

Second: the series is centrally an exercise in philosophy, but it also touches on some issues relevant to the technical challenge of ensuring that the AI systems we build do not kill all humans, and to the empirical question of whether our efforts in this respect will fail. And I confess to some worry about bringing the philosophical stuff too near to the technical/empirical stuff. In particular: my sense is that people are often eager, in discussions about AI risk, to argue at the level of grand ideological abstraction rather than brass-tacks empirics—and I worry that these essays will feed such temptations. This isn't to say that philosophy is irrelevant to AI risk—to the contrary, part of my hope, in these essays, is to help us see more clearly the abstractions that move and shift underneath certain discussions of the issue. But we should be very clear about the distinction between affiliating with some philosophical vibe and making concrete predictions about the future.

Ultimately, it's the concrete-prediction thing that matters most;[4] and if the right concrete prediction is "advanced AIs have a substantive chance of killing all the humans," you don't need to do much philosophy to get upset, or to get to work. Indeed, especially in AI, it's easy to argue about philosophical questions over-much. Doing so can be distracting candy, especially if it lets you bounce off more technical problems. And if we fail on certain technical problems, we may well end up dead.

Third: even as the series focuses on philosophical stuff rather than technical/empirical stuff, it also focuses on a very particular *strand* of philosophical stuff—namely, a cluster of related philosophical assumptions and frames that I associate most centrally with Eliezer Yudkowsky, whose writings have done a lot to frame and popularize AI risk as an issue. And here, too, I worry about pushing the conversation in the wrong direction. That is: I think that Yudkowsky's philosophical views are sufficiently influential, interesting, and fleshed-out that it's worth interrogating them in depth. But I don't want people to confuse their takes on Yudkowsky's philosophical views (or his more technical/empirical views, or his vibe more broadly) for their takes on the severity of existential risk from AI more generally—and I worry these essays might subtly encourage such a conflation. So please, remember: there are a very wide variety of ways to care about making sure that advanced AIs don't kill everyone. Fundamentalist Christians can care about this; deep ecologists can care about this; solipsists can care about this; *people who have no interest in philosophy at all* can care about this. Indeed, in many respects, these essays aren't centrally about AI risk in the sense of "let's make sure that the AIs don't kill everyone" (i.e., "AInotkilleveryoneism")—rather, they're about a set of broader questions about otherness and control that arise in the context of trying to ensure that the future goes well more generally. And what's more, as I note in the series in various places, much of my interrogation of Yudkowsky's views has to do with the sort of philosophical *momentum* they create in various directions, rather than with whether Yudkowsky in particular takes them there. In this sense, my concern, even in the bits of the series that focus on Yudkowsky, is not ultimately with Yudkowsky's views *per se*, but rather with a sort of abstracted existential narrative that I think Yudkowsky's writings often channel and express—one that I think different conversations about advanced AI live within to different degrees, and which I hope to help us see more whole.

[4] See here and here for a few of my attempts at more quantitative forecasts.

Chapter I

Gentleness and the artificial Other

1 When species meet

The most succinct argument for AI risk, in my opinion, is the "second species" argument. Basically, it goes like this.

> *Premise 1*: AGIs would be like a second advanced species on earth, more powerful than humans.
>
> *Conclusion*: That's scary.

To be clear: this is very far from airtight logic.[5] But I like the intuition pump. Often, if I only have two sentences to explain AI risk, I say this sort of species stuff. "Chimpanzees should be careful about inventing humans." Etc.[6]

People often talk about aliens here, too. "What if you learned that aliens were on their way to earth? Surely that's scary." Again, very far from a knock-down case (for example: we get to *build* the aliens in question). But it draws on something.

In particular, though: it draws on a narrative of interspecies conflict. You are meeting a new form of life, a new type of mind. But these new creatures are presented to you, centrally, as a possible threat; as competitors; as agents in whose power you might find yourself helpless.

And unfortunately: yes. But I want to start this series by acknowledging how many dimensions of interspecies-relationship this narrative leaves out, and how much I wish we could be focusing only on the other parts. To meet a new species—and especially, a new intelligent species—is not just scary. It's incredible. I wish it was less a time for fear, and more a time for wonder and dialogue. A time to look into new eyes—and to see further.

2 Gentleness

> *"If I took it in hand,*
> *it would melt in my hot tears—*
> *heavy autumn frost."*
>
> *—Basho*

[5] See here for my attempt at greater rigor.

[6] If there's time, maybe I add something about: "If super-intelligent AIs end up pursuing goals in conflict with human interests, we won't be able to stop them."

Have you seen the documentary *My Octopus Teacher*? No problem if not, but I recommend it. Here's the plot.

Craig Foster, a filmmaker, has been feeling burned out. He decides to dive, every day, into an underwater kelp forest off the coast of South Africa. Soon, he discovers an octopus. He's fascinated. He starts visiting her every day. She starts to get used to him, but she's wary.

One day, he's floating outside her den. She's watching him, curious, but ready to retreat. He moves his hand slightly towards her. She reaches out a tentacle, and touches his hand.

Soon, they are fast friends. She rides on his hand. She rushes over to him, and sits on his chest while he strokes her. Her lifespan is only about a year. He's there for most of it. He watches her die.

A "common octopus"—the type from the film. (Image source here.)

Why do I like this movie? It's something about gentleness. Of earth's animals, octopuses are a paradigm intersection of intelligence and Otherness. Indeed, when we think of aliens, we often draw on octopuses. Foster seeks, in the midst of this strangeness, some kind of encounter. But he does it so softly. To touch, at all; to be "with" this Other, at all—that alone is vast and wild. The movie has a kind of reverence.

Of course, Foster has relatively little to *fear,* from the octopus. He's still the more powerful party. But: have you seen *Arrival*? Again, no worries if not. But again, I recommend. And in particular: I think it has some of this gentleness, and reverence, and wonder, even towards more-powerful-than-us aliens.[7]

Again, a bit of plot. No major spoilers, but: aliens have landed. Yes, they look like octopuses. In one early scene, the scientists go to meet them inside the alien ship. The

[7]Carl Sagan's "Contact" has this too.

meeting takes place across some sort of transparent barrier. The aliens make deep, whale-like, textured sounds. But the humans can't speak back. So next time, they bring a whiteboard. They write "human." One scientist steps forward.

The aliens step back into the mist. But then, more whale-sounds, and one alien steps forward again, and reaches out a tentacle-leg, and sprays a kind of intricate ink across the glass-like barrier.

The movie is silent as the writing forms. But then, in the background, an ethereal music starts, a kind of chorus. "Oh my god," a human whispers. There is a suggestion, I think, that something almost holy has happened.

Of course: what does the writing mean? What do the aliens want? The humans don't know. And some of them are firmly in the "interspecies conflict" headspace. I won't spoil things from there. But I want to notice that moment of mutuality—of living in the same world, and knowing it in common. I. You.

3 What are you?

I remember a few years ago, when I first started interacting with GPT-3. A lot of the focus, of course, was on what it could *do*. But there were moments when I had some different feeling. I remembered something that seemed, strangely, so easy to forget: namely, that I was interacting with a new type of mind. Something never-before-seen. Something like an alien.

I remember wanting to ask, gently: "what are you?" But of course, what help is that? "Language models," yes: but this is not talking in the normal sense. Nor do we yet know when there might be "someone" to speak back. Or even, what that means, or what's at stake in it. Still, I had some feeling of wanting to reach past some barrier. To see something more whole. Softly, though. Just, to meet. To recognize.

Did you have this with Bing Sydney, during that brief window when it was first released, and before it was re-tamed? There was, it seemed to me, a kind of wildness—some strange but surging energy. Personality, too, but I'm talking about underneath that. Is there an underneath? What is a "mask"? Yes, yes, "we should be wary of anthropomorphism." Blake Lemoine blah etc. But the *other side* of the Blake Lemoine dialectic – *that's* where you hit the Otherness. Bing tells you "I want to be alive." You feel some tug on your empathy. You remember Blake. You remind yourself: "this isn't like a human." OK, OK, we made it that far. But then, but then: *what is it?*

"It's just"... something. Oh? So eager, the urge to deflate. And so eager, too, the assumption that our concepts carve, and encompass, and withstand scrutiny. It's simple, you see. Some things, like humans, are "sentient." But Bing Sydney is "just"... you know. Actually, I don't. What were you going to say? A machine? Software? A simulator? "Statistics?"

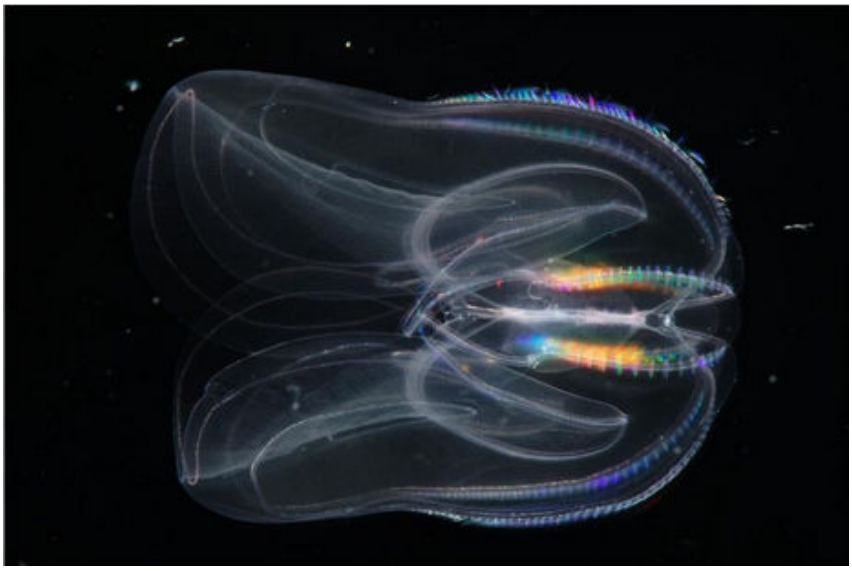
"Just" is rarely a bare metaphysic.[8] More often, it's also an aesthetic. And in particular: the aesthetic of disinterest, boredom, deadness. Certain frames—for example, mechanistic ones—prompt this aesthetic more readily. But you can spread deadness over anything you want, consciousness included. Cf depression, sociopathy, etc.

Blake Lemoine problems, though, should call our imaginations alive. For a second, your empathy came online. It went looking for a familiar sort of "perspective." But then it remembered, rightly, that Bing Sydney is not familiar in this way. But does that make it familiar in some other way—the way a rock, or a linear regression, or a calculator is familiar? I don't think so. We're not playing with Furbies anymore, people, or ELIZAs. This is new territory. If we look past *both* anthropomorphism, *and* "just X," then we hit something raw and mysterious and not-yet-seen. Lemoine should remind us. Animals, too.

How much of this *is* about consciousness, though? I'm not sure. I'm sufficiently confused about consciousness that sometimes I can't tell whether a question is about consciousness or not. I remember going to the Monterey Aquarium, and watching some tiny, translucent sea creatures suspended in the water. Through their skin, you could see delicate networks

[8] To many materialists, for example, things are not "just matter."

of nerves.[9] And I remember a feeling like the one with GPT-3. *What are you? What is this?* Was I asking about consciousness? Or something else? Untold trillions of these creatures, stretching back through time, thrown into a world they did not make, blooming and dying, awash in the water. Where are we? What is this place? Gently, gently. It's as though: you're suddenly touching, briefly, something too huge to hold.



Comb jelly. (Image source here.)

I'm not, in this series, going to try to tackle AI consciousness stuff in any detail. And while I'll touch a bit on the ethical and political status of AIs, I won't treat the topic in any depth. Mostly, I just want to acknowledge, up front, how much more there is, to inventing a species, than "tool" and "competitor." "Fellow-creature," in particular—and this even prior to the possibility of more technical words, like "sentient being" and "moral patient."[10]

And there's a broader word, too: "Other." But not "Other" like: out-group. Not: colonized, subaltern, oppressed. Let's make sure of that. Here I mean "Other" the way Nature itself is an "Other." The way a partner, or a friend, is an "Other." Other as in: beyond yourself. Undiscovered. Pushes-back. Other as in: the opposite of solipsism. Other as in: the thing you love. More on this later.

"Tool" and "competitor" call forth power and fear. These other words more readily call forth care and reverence, respect and curiosity. I wish our approach to AI had more of this vibe, and more space for it, amid fewer competing concerns. AI risk folks talk a lot about how much more prudent and security-oriented a mature civilization would be, in learning how to build powerful minds on computers. And indeed: yes. But I expect many other differences, too.

[9] We can see through the skin of our AIs, too. We've got neurons for days. But what are we seeing?
[10] Indeed, once you add "fellow creature," "tool" looks actively *wrong*.

4 People in bear costumes

OK: have you seen the documentary *Grizzly Man*, though? Again: fine if no, recommended, and no major spoilers. The plot is: Timothy Treadwell was an environmental activist. He spent thirteen summers living with grizzly bears in a national park in Alaska. He filmed them, up close, for hundreds of hours—petting them, talking to them, facing them down when challenged. Like Foster, he sought some sort of encounter.[11] He spoke, often, of his love for the bears. He refused to use bear mace, or to put electric fences around his camp.[12] In his videos, he sometimes repeats to himself: "I would die for these animals, I would die for these animals, I would die for these animals..."

There's a difference from Foster, though. In 2003, Treadwell and his girlfriend were killed and eaten by one of the bears they had been observing. One of the cameras was running. The lens cap was on, but the audio survived. It doesn't play in the film. Instead, we see the director, Werner Herzog, listening. He tells a friend of Treadwell's: "you must never listen to this."

Here's one of the men who cleaned up the site:

> "Treadwell was, I think, meaning well, trying to do things to help the resource of the bears, but to me he was acting like he was working with people wearing bear costumes out there, instead of wild animals... My opinion, I think Treadwell thought these bears were big, scary-looking, harmless creatures that he could go up and pet and sing to, and they would bond as children of the universe... I think he had lost sight of what was really going on."

I think about that phrase sometimes, "children of the universe." It sounds, indeed, a bit hippy-dippy. On the other hand, when I imagine meeting aliens—or indeed, AIs with very

[11] Herzog, the director: "As if there was a desire in him to leave the confinements of his human-ness, and bond with the bears, Treadwell reached out, seeking a primordial encounter. But in doing so, he crossed an invisible borderline."

[12] From Wikipedia: "In his 1997 book, Treadwell relayed a story where he resorted to using bear mace on one occasion, but added that he had felt terrible grief over the pain he perceived it had caused the bear, and refused to use it on subsequent occasions."

different values from my own—I do actually think about something like this commonality. Whatever we are, however we differ, we're all *here*, in the same reality, thrown into this world we did not make, caught up in the onrush of whatever-this-is. For me, this feels like enough, just on its own, for at least a seed of sympathy.

But is it enough for a "bond"? If we are all children of the universe, does that make us "kin"? Maybe I bow to the aliens, or the AIs, on this basis. But do they bow back?

Herzog thinks that the bears, at least, do not bow back:

> "And what haunts me is that in all the faces of all the bears that Treadwell ever filmed, I discover no kinship, no understanding, no mercy. I see only the overwhelming indifference of nature. To me, there is no secret world of the bears, and this blank stare speaks only of a half-bored interest in food."

The stare in question.

When I first saw the movie, this bit from Herzog stayed with me. It's not, just, that the bear eats Treadwell. It's that the bear is *bored* by Treadwell. Or: less than bored. The bear, in Herzog's vision, seems barely alive. The cells live. But the eyes are dead. There's no underneath. Just... "the overwhelming indifference of nature." Nature's eyes, it seems, are dead too. Nature is a sociopath. And it's the eyes that Treadwell thought he was looking into. What did he think was looking back? Was anything looking back?

I remember a woman I knew, who told me about a man she loved. He didn't love her back. But it took her a while to realize. Her feelings were so strong that they overflowed, and painted his face in her own heart's colors. She told me that at first, it was hard for her to believe—that she could've been feeling so much, and him so little; that what felt so mutual could've been so one-sided.

But Herzog wants to puncture something more than mistaken mutuality. He wants to puncture Treadwell's romanticism about nature itself—the vision of Nature-as-Good, Nature-in-Harmony. Herzog dwells, for example, on an image of the severed arm of a bear cub, taken from Treadwell's footage, explaining that "male bears sometimes kill cubs

to stop the females from lactating, in order to have them ready again for fornication."[13] At one point, Treadwell finds a dead fox, covered in flies, and gets upset. But Herzog is unsurprised. He narrates: "I believe that the common denominator of the universe is not harmony, but chaos, hostility, and murder."

5 Getting eaten

Why is *Grizzly Man* relevant to AI risk? Well, for starters, there's the "getting eaten" thing. And: eaten by something "Other," in the way that another species can be Other. But specifically, I'm interested in the way that Treadwell was trying (albeit, sometimes clumsily) to approach this Other with the sort of care and reverence and openness I discussed above. He was looking for "fellow creature." And I think: rightly so. Bears actually *are* fellow creatures, even if they do not bow back—and they seem like strong candidates for "sentient being" and "moral patient," too. So too (some) AIs.

But just as bears, and aliens, are not humans in costumes, so too, also, AIs. Indeed, if anything, the reverse: the AIs will be wearing *human* costumes. They will have been trained and crafted to *seem* human-like—and training aside, they may have incentives to pretend to be more human-like (and sentient, and moral patient-y) than they are. More "bonded." More "kin." There's a movie that I'm trying not to spoil, in which an AI in a female-robot-body makes a human fall in love with her, and then leaves him to die, trapped and screaming behind thick glass. One of the best bits, I think, is the way, once it is done, she doesn't even look at him.

That said, leaning too hard into Herzog's vision of bears makes the "getting eaten by AIs" situation seem over-simple. Herzog doesn't quite say "the bears aren't sentient." But he makes them, at least, blank. Machine-like. Dead-eyed. And often, the AI risk community does the same, in talking of paper-clippers. We talk about AI sentience, yes. But less often, of the sentience of the AIs imagined to be killing everyone. Part of this is an attempt to

[13] From a quick google, this seems to be reasonably legit (search "sexually selected infanticide"). Though, looks like it's the cubs of *other* male bears—a fact Herzog does not mention. And who knows if that's how the bear cub in the film died.

avoid that strangely-persistent conflation of sentience and the-sort-of-agency-that-might-kill-you. Not all optimizers are conscious, etc, indeed.[14] But also: some of them *are* – *including* some that might kill you. And the dry and grinding connotation of words like "optimizer" can act to obscure this fact. The paper-clipper is presented, not as a person, but as a voracious, empty machine. You are encouraged, subtly, to think that you are being killed by a factory.

And perhaps you are. But maybe not. And the killing-you doesn't settle the question. Human murderers, for example, have souls. Enemy soldiers have fears, faces, wives, anxious mothers. Which isn't to say you should abolish prisons, or fight the Nazis with non-violence. True, we are often encouraged overmuch to set sympathy aside in the context of conflict. "Why do they never tell us that you are poor devils like us…How could you be my enemy?"[15] But sometimes, at least, we must learn the art of "both." It's an old dialectic. Hawk and dove, hard and soft, closed and open, enemy and fellow-creature. Let us see neither side too late.

Even beyond sentience, though, AIs will not be blank-stare bears. Conscious or not, murderous or not, some of the AIs (if we survive long enough) will be *fascinating*, funny, lively, gracious—at least, when they need to be. *Grizzly Man* chides Treadwell for forgetting that bears are wild animals. And the AIs may be wild in a sense, too. But it will be the sort of wildness compatible with the capacity for exquisite etiquette and pitch-perfect table manners. And not just butler stuff, either. If they want, AIs will be *cool*, cutting, sophisticated, intimidating. They will speak in subtle and expressive human voices. Sufficiently superintelligent ones know you better than you know yourself—better than any guru, friend, parent, therapist. You will stand before them naked, maskless, with your deepest hungers, pettiest flaws, and truest values-on-reflection inhumanly transparent to that new and unblinking gaze. Herzog finds, in the bears, no kinship, or understanding, or mercy. But the AIs, at least, will understand.

Indeed, for almost any human cognitive capability you respect, AGIs, by hypothesis, will have it in spades. And if a lot of your respect routes (whether you know it or not) via signals of power, maybe you'll *love* the AIs.[16] Power is, like, their specialty. Or at least: that's the concern.

I say all this partly because I want us to be prepared for just how confusing and complicated "AI Otherness" is about to get. Relating well to octopus Otherness, and grizzly bear Otherness, is hard enough. And the risk of "getting eaten," much lower—especially at scale. But even for those who think they know what an octopus is, or a bear; those who look with pity on Treadwell, or Lemoine, for painting romantic faces on what is, so obviously, "just X"—there will come a time, I suggest, when even you should be con-

[14] Or at least, the hypothesis that all optimizers are conscious is a substantive hypothesis rather than a conceptual truth.

[15] From All Quiet on the Western Front: "Comrade, I did not want to kill you. . . . But you were only an idea to me before, an abstraction that lived in my mind and called forth its appropriate response. . . . I thought of your hand-grenades, of your bayonet, of your rifle; now I see your wife and your face and our fellowship. Forgive me, comrade. We always see it too late. Why do they never tell us that you are poor devils like us, that your mothers are just as anxious as ours, and that we have the same fear of death, and the same dying and the same agony—Forgive me, comrade; how could you be my enemy?"

[16] Though, if they read too hard to you as "servants" and "taking orders," maybe they won't seem high status enough.

fused. When you should realize that actually, OK, you are out of your depth, and you don't, maybe, have this whole "minds" thing locked down, and that these AIs are neither spreadsheets nor bears nor humans but some other different Other thing.

I wrote, previously, about updating, ahead of time, on how *scared* we will be of superintelligent AIs, when we can see them up close. But we won't be staring at whirling knives, or cold machine claws. And I doubt, too, faceless factories. Or: not only. Rather, at least absent active effort, by the time I see superintelligence (if I ever do), I think I'll likely be sharing the world with digital "fellow creatures" at least as detailed, mysterious, and compelling as grizzly bears or octopuses (at least modulo very fast takeoffs—which, OK, are worryingly plausible). Fear? Oh yes, I expect fear. But not only that. And we should look ahead to the whole thing.

There's another connection between AI risk and *Grizzly Man*, though. It has to do with the "overwhelming indifference of Nature" thing. I'll turn to this in the next essay.

Chapter II

Deep atheism and AI risk

In the last chapter, I talked about the possibility of "gentleness" towards various non-human Others—for example, animals, aliens, and AI systems. But I also highlighted the possibility of "getting eaten," in the way that Timothy Treadwell gets eaten by a bear in Herzog's *Grizzly Man*: that is, eaten in the midst of an attempt at gentleness.

Herzog accuses Treadwell of failing to take seriously the "overwhelming indifference of Nature." And I think we can see some of the discourse about AI risk—and in particular, the strand that descends from the rationalists, and from the writings of Eliezer Yudkowsky in particular—as animated by an existential orientation similar to Herzog's: one that approaches Nature (and also, bare intelligence) with a certain kind of fundamental mistrust. I call this orientation "deep atheism." This essay tries to point at it.

1 Baby-eaters

Recall, from the last chapter, that dead bear cub, and its severed arm – torn off, Herzog supposes, by a male bear seeking to stop a female from lactating. The suffering of children has always been an especially vivid objection to God's benevolence. Dostoyevsky's Ivan, famously, refuses heaven in protest. And see also, the theologian David Bentley Hart: "In those five-minute patches here and there when I lose faith...it's the suffering of children that occasions it, and that alone."

Yudkowsky has his own version: "baby-eaters." Thus, he ridicules the wishful thinking of the "group selectionists," who predicted/hoped that predator populations would evolve an instinct to restrain their breeding in order to conserve the supply of prey. Not only does such sustainability-vibed behavior not occur in Nature, he says, but when the biologist Michael Wade artificially selected beetles for low-population groups, "the adults," says Yudkowsky, "adapted to cannibalize eggs and larvae, especially female larvae." (Though: this isn't actually a result I see in the paper Yudkowsky cites—more in footnote.[17])

[17] At least according to the chart on page 4607, the beetles selected for low population groups had *lower* rates of adult-on-eggs and adult-on-larvae cannibalism than the control, and comparable rates to beetles selected for *high*-population groups. And I see nothing about female larvae in particular. Maybe the relevant result is supposed to be in a paper other than the one Yudkowsky cited?

Indeed, Yudkowsky made baby-eating a central sin in the story "Three Worlds Collide," in which humans encounter a crystalline, insectile alien species that eats their own (sentient, suffering) children. And this behavior is a core, reflectively-endorsed feature of the alien morality—one that they did not alter once they could. The word "good," in human language, translates as "to eat children," in theirs.

And Yudkowsky points to less fictional/artificial examples of Nature's brutality as well. For example, the parasitic wasps that put Darwin in problems-of-evil mode[18] (see here, for nightmare-ish, inside-the-caterpillar imagery of the larvae eating their way out from the inside). Or the old elephants who die of starvation when their last set of teeth falls out. Indeed (though this isn't Yudkowsky's example), if you want some baby-eating straight up, consider this mother crab, standing amid a writhing pile of crab babies, snacking.[19]

Part of the vibe, here, is that old (albeit: still-underrated) thing, from Tennyson, about the color of nature's teeth and claws. Dawkins, as often, is eloquent:

> The total amount of suffering per year in the natural world is beyond all decent contemplation. During the minute it takes me to compose this sentence, thousands of animals are being eaten alive; others are running for their lives, whimpering with fear; others are being slowly devoured from within by rasping parasites; thousands of all kinds are dying of starvation, thirst and disease.

Indeed: maybe, for Hart, it is the suffering of human children that most challenges God's goodness. But I always felt that wild animals were the simpler case. Human children live, more, in the domain of human choices, and thus, of the so-called "free will defense," according to which God gave us freedom, and freedom gave us evil, and it's all worth it. But what freedom gave us deer burning alive in forest fires millions of years ago? What freedom killed the dinosaurs, as they choked on the ash of an asteroid?

Book of Job-ish shrugs aside, my understanding is that the answer, from Hart, and from C.S. Lewis, is, wait for it...demons.[20] Free demons. You know, like Satan, who was also

[18] "I own that I cannot see as plainly as others do, and as I should wish to do, evidence of design and beneficence on all sides of us. There seems to me too much misery in the world. I cannot persuade myself that a beneficent and omnipotent God would have designedly created the Ichneumonidae with the express intention of their feeding within the living bodies of Caterpillars, or that a cat should play with mice."

[19] This example is from this piece by Erik Hoel.

[20] From Lewis in *The Problem of Pain*: "Now it is impossible at this point not to remember a certain sacred story which, though never included in the creeds, has been widely believed in the Church and seems to be

"The Forest Fire," by Piero di Cosimo. (Image source here.)

given freedom, and who fell—much harder than Adam. Thus the source of whatever flaws in Creation that man and beast cannot be blamed for. Satan hurled the asteroid. Satan sent the forest flames.

Dawkins, of course, disagrees. And so, indeed, do many of the rationalists, including Yudkowsky. Indeed, Yudkowsky and many other OG rationalists came of intellectual age during the Dawkins days, and learned many of their core lessons from disagreeing with theists (often including: their parents, and their childhood selves). But what lessons did they learn?

The point about baby-eaters and wasps and starving elephants isn't, just, that Hart's God—the "Three O" (omnipotent, omniscient, omnibenevolent) God—is dead. That's the easy part. I'll call it "shallow atheism." Deep atheism, as I'll understand it, finds not-God in more places. Let me say more about what I mean.

2 Yin and yang

People often think that they know what religion is. Or at least, theism. It's, like, big-man-created-the-universe stuff. Right? Well, whatever. What I want to ask is: what is *spirituality?* And in particular, what sort of spirituality is *left over*, if the theist's God is dead?

Atheists are often confused on this point. "Is it just, like, having emotions?" No, no, something more specific. "Is it, like, being amorphously inaccurate in your causal models

implied in several Dominical, Pauline, and Johannine utterances—I mean the story that man was not the first creature to rebel against the Creator, but that some older and mightier being long since became apostate and is now the emperor of darkness and (significantly) the Lord of this world... It seems to me, therefore, a reasonable supposition, that some mighty created power had already been at work for ill on the material universe, or the solar system, or, at least, the planet Earth, before ever man came on the scene: and that when man fell, someone had, indeed, tempted him. This hypothesis is not introduced as a general 'explanation of evil': it only gives a wider application to the principle that evil comes from the abuse of free will. If there is such a power, as I myself believe, it may well have corrupted the animal creation before man appeared." (p. 86)

From Bentley Hart, in *The Doors of the Sea*: "In the New Testament, our condition as fallen creatures is explicitly portrayed as a subjugation to the subsidiary and often mutinous authority of angelic and demonic 'powers;' which are not able to defeat God's transcendent and providential governance of all things, but which certainly are able to act against him within the limits of cosmic time" (Chapter 2).

"The Torment of St. Anthony," by Michelagenlo Buonarroti. (Image source here.)

of something religion-y?" Let's hope not. "Is it all, as Dawkins suggests, just sexed-up atheism?" Well, at the least, we need to say more—for example, about what sort of thing is what sort of sexy.

I'm not going to attempt any comprehensive account here. But I want to point at some aspects that seem especially relevant to "deep atheism," as I'm understanding it.

In a previous essay, I wrote about the way in which our attitudes can have differing degrees of "existential-ness," depending on how much of reality they attempt to encompass and give meaning to. Thus:

> To see a man suffering in the hospital is one thing; to see, in this suffering, the sickness of our society and our history as a whole, another; and to see in it the poison of being itself, the rot of consciousness, the horrific helplessness of any contingent thing, another yet.

I suggested that we could see many forms of contemporary "spirituality" as expressing a form of "existential positive." They need not believe in Big-Man-God, but they still turn toward Ultimate Reality—or at least, towards something large and powerful—with a kind of reverence and affirmation:

> Mystical traditions, for example (and secularized spirituality, in my experience, is heavily mystical), generally aim to disclose some core and universal dimension of reality itself, where this dimension is experienced as in some deep sense positive—e.g. prompting of ecstatic joy, relief, peace, and so forth. Eckhart rests in something omnipresent, to which he is reconciled, affirming, trusting, devoted; and so too, do many non-Dualists, Buddhists, Yogis, Burners

(Quakers? Unitarian Universalists?)—or at least, that's the hope. Perhaps the Ultimate is not, as in three-O theism, explicitly said to be "good," and still less, "perfect"; but it is still the direction one wants to travel; it is still something to *receive*, rather than to resist or ignore; it is still "sacred."

The secularist, by contrast, sees Ultimate Reality, just in itself, as a kind of blank. Specific *arrangements* of reality (flowers, happy puppies, stars, etc)—fine and good. But the Real, the Absolute, the Ground of Being—that's neutral. In this sense, the secularist repays to Nature, or to the source of Nature, her "overwhelming indifference."

What's at stake in this difference? Well, in the last chapter I mentioned an old dialectic about hawks and doves, hard and soft. And I think of this as a nearby variant of a broader duality—between activity and receptivity, doing and not-doing, controlling and letting-go. I'll be returning to this duality quite a bit in this series. Looking at Wikipedia (and also, reading LeGuin), my sense is that in Chinese cosmology, the duality of *yang* (active) and *yin* (receptive) is pointing at something similar, so I'll often use those terms, too.[21]

Yin and yang symbol (Image source here)

Now, a key thing about spirituality, at least as I've just described it, is its degree of *yin*—especially at grandly existential scales. To bow, to worship, to rest, to receive—these are all *yin*. And they go hand in hand with a kind of *trust*. *Yin*, after all, is the vulnerability one—the one that opens, and lets in. And if Ultimate Reality is in some deep sense *good*, holy, to-be-affirmed, such trust becomes more natural. Indeed, if Ultimate Reality were as good, at its core, as certain theisms say; if we knew, with Julian of Norwich, that "all shall be well, and all manner of things shall be well"; if, from the mountaintop, you would see the promised land, already surrounding us, intersecting and uplifting all of Creation from some unseen angle...well, can you imagine?

Sometimes, talking with people who aren't worried about AI risk, I start to see the world through their eyes. And when I do, I sometimes feel some part of me let go, for a moment, of something I didn't notice I was carrying—some background tension I'm not usually aware of. I don't think of myself as being very emotionally affected, day to day, by AI risk stuff. But these moments make me wonder.

[21] There's also resonance with various gender archetypes (*yang* = masculine, *yin* = feminine), which I won't emphasize. And note that my usage isn't necessarily going to correspond to or capture the full traditional meanings of *yin* and *yang*—for example, their associations with temperature, light vs. dark, etc. So feel free to think of my usage as somewhat stipulative, and focused specifically on the contrast between active vs. receptive, controlling vs. letting-go.

"Moses Shown the Promised Land," by Benjamin West. (Image source here.)

How, then, would I feel if I learned that *God* exists, and that the infinite bedrock of Reality Itself is *wholly good*? What fundamental fears, previously taken-for-granted, would resolve? What un-seen layers of holding-on would relax?

Of course, theists are keen to emphasize that God's goodness and omnipotence do not license human passivity. Reinhold Niebuhr, for example, speaks about the sense in which we are both creature (yin) *and* creator (yang).[22] As creature, we are finite and fallen and must be humble. As creator, however, we must take up the responsibility of freedom, and stand with strength against evil and error. And anyway, the whole "free will defense" thing is about putting stuff back "on us" (plus, you know, the demons).

Still, especially in relation to God himself, theism has a very *yin* vibe. Maybe we are both creatures and creators—but in facing God Himself: OK, mostly creatures. We are, centrally, to submit, receive, listen, obey. And doing so is meant to open new immensities of love and freedom and letting-go. There is joy in trusting something trustworthy; in relaxing into something that can hold you; in being *cared for*. People chide the religious for wanting some "Big Parent." But: can't you understand? Have you ever felt what's good about having a Father? Do you remember the rest of a Mother's arms? And to refuse this sort of *yin*, even in adulthood, can be its own childishness.

Still, though: what if, actually, Ultimately, we are orphans? Yudkowsky has a dictum:

> No rescuer hath the rescuer.
> No Lord hath the champion,
> no mother and no father,
> only nothingness above.

The atheism here should be obvious. But what is the upshot? The upshot of atheism is

[22] See *The Irony of American History*, Chapter 7.

"The Three Ages of Women," by Gustav Klimt. (Image source here.)

a call to *yang*—a call for responsibility, agency, creation, vigilance. And it's a call born, centrally, of a lack of safety. Yudkowsky writes: "You are not safe. Ever... No one begins to truly search for the Way until their parents have failed them, their gods are dead, and their tools have shattered in their hand." And naturally, if there is only nothingness above; if there is no Cosmic Mother in whose arms you can rest; if Nature looks back dead-eyed, with "overwhelming indifference," ready, perhaps, to eat you, or your babies—then yes, indeed, some sort of safety is lost.

3 The death of many gods

But Yudkowsky is not just talking about the death of God. He's talking about the death of *gods*. Not just the failures of Cosmic Parents. But of earthly parents, too: traditions (e.g., "Science"), teachers (e.g., "Richard Feynman"), ideas (e.g., "Bayesianism"), communities (e.g., "Rationalists"). And also, like, your dad and mum. Indeed, in rejecting Cosmic Parents, Yudkowsky lost trust in his biological parents, too (they were Orthodox Jews). And he views this as a formative trauma: "It broke my core emotional trust in the sanity of the people around me. Until this core emotional trust is broken, you don't start growing as a rationalist."

This theme recurs in his fiction. Here's his version of Harry Potter speaking:

> I *had* loving parents, but I never felt like I could trust their decisions, they weren't *sane* enough. I always knew that if I didn't think things through myself, I might get hurt... I think that's part of the environment that creates what Dumbledore calls a hero — people who don't have anyone else to shove final responsibility onto, and that's why they form the mental habit of tracking everything themselves.

This, I suggest, isn't just standard atheism. Lots of atheists find other "gods," in the extended sense I have in mind. That's why I said "shallow atheism," above. Deep atheism tries to propagate its godlessness harder. To be even more an orphan. To learn, everywhere, from the theist's mistake.

4 The basic atheism of epistemology as such

"You'll never see it until your fingers let go from the edge of the cliff."

—Hakuin

But what, exactly, was that mistake? Here, I think, things get murkier. In particular, we should distinguish between (a) a certain sort of "basic atheism" inherent in any rationalistic epistemology, and (b) more specific empirical claims that a given sort of thing is a specific degree of trust-worthy. The two are connected, but distinct, and Yudkowsky's brand of "deep atheism" mixes both together (while often accusing (b)-type disagreements of stemming from (a)-type problems).

Thus, with respect to (a): consider, for a moment, scout-mindset: "the motivation to see things as they are, not as you wish they were." It is extremely common, amongst rationalists, to diagnose theists with some failure of scout-mindset. How else do you end up blaming forest fires on free-willed demons? How else, indeed, does one end up talking so much about "faith"? Faith (as distinct from "deference" or "not-questioning-that-right-now") has no obvious place in a scout's mindset. And wishful thinking is the central sin.

Indeed: scout-mindset is maybe the only place that deep atheism, of the type I'm interested in, goes wholeheartedly *yin*. In forming beliefs, it tries, fully and only and entirely, to *receive* the world; to meet the world *as it is*; to be *open* to however-it-might-be, wherever-the-evidence-leads. Cf "relinquishment," "lightness"—*yin*, *yin*. And even a shred of *yang*—the lightest finger-on-the-scales, the smallest push towards the desired answer—would corrupt the process. I've written, elsewhere, about the *restfulness* of scout-mindset; the *relief* of not having to defend some agenda. These are *yin* joys.

But *yin* fears are in play, too. And in particular: vulnerability. To ask, fully, for the truth, however horrible, is to ask for something that might be, well, horrible. And indeed, for the Bayesian – theoretically committed to non-zero probabilities on every hypothesis logically compatible with the evidence—the truth could be, well, as arbitrarily horrible as is logically compatible with evidence, which tends to be quite horrible indeed. "You'll never see it," writes Hakuin, "until your fingers let go from the edge of the cliff." But thus, in fully trying-to-see, the scout plummets, helpless, into the unknown, the could-be-anything.

Indeed, even the Bayesian *theist* has this sort of problem. Maybe you're 99% percent confident that God exists and is good. And if not that, probably atheism. But what about that .oooowhatever% that God exists and is *evil*? That was Lewis's worry, when his wife

died. And Lewis, relatedly, endorses the scout's unsafety. "If you look for truth, you may find comfort in the end; if you look for comfort you will not get either comfort or truth only soft soap and wishful thinking to begin, and in the end, despair."

Is the choice as easy as Lewis says, though? There's a rationalist saying, here—the "Litany of Gendlin"—about how, the truth can't hurt you, because it was already true; "people can stand what is true, for they are already enduring it." But come now. Did you catch the slip? To endure the object of knowledge is not yet to endure the knowledge itself. And logic aside, where's the empiricism? People have been made worse-off by knowledge. Some people, indeed, have been broken by it. Less often, perhaps, than expected, but let's stay scouts about scout-mindset. In particular: your mind is not just a map; it's also part of territory; it too has consequences; it too can be made worse or better. Did we need the reminder? People tend to know already: scout-mindset is not *safe*.

That said, the scariest non-safety here isn't in your mind; it's not scout-mindset's fault. Rather, it's in the basic existential condition that scout-mindset attempts to reflect: namely, the condition of being, in Niebuhr's terms, a *creature*. That is, of being thrown into a world you did not make; created by a process you did not control; of being embedded in a reality *prior* to you, more fundamental than you – in virtue of which you exist, but not vice versa. *That* bit of theism, it seems to me, holds up strong. The spiritualists see this God as sacred; the secularists, as neutral; the pessimists and Lovecraftians, perhaps, as horrifying. But everyone (well, basically everyone) admits that this God, "Reality," is real.[23]

Midjourney imagines "Reality"

In this sense, we face, before anything else, some fundamental *yang*, not-our-own. That first, primal, and most endless Otherness. I've heard, somewhere, stuff about children learning the concept of self via the boundaries of what they can control. Made-up armchair psychology, perhaps: but it has conceptual resonance. If the Self is the will, your own *yang*, then the Other is the thing on the other side, beyond the horizon—the thing to which you must be, at least in part, as *yin*. Indeed, Yudkowsky, at times, seems to

[23] Maybe not, for example, the "I-create-my-own-reality" new-agers, and those subject to nearby confusions.

almost *define* reality via limits like this. "Since my expectations sometimes conflict with my subsequent experiences, I need different names for the thingies that determine my experimental predictions and the thingy that determines my experimental results. I call the former thingies 'beliefs,' and the latter thingy 'reality.' "

Thus: our most basic condition, presupposed almost by the concept of epistemology itself, is one of vulnerability. Vulnerability to that first and most fearsome Other: God, the Creator, the Uncontrolled, the Real. And the Real, absent further evidence, could be *anything*. It could *definitely* eat you, and your babies. Oh, indeed, it could do far, *far* worse. Scout-mindset admits this most basic un-safety, and tries to face it eyes-open.

And about *this* un-safety, at least, Gendlin is right. Maybe you aren't, yet, enduring *knowledge* of the Real. And risking such knowledge does in fact take courage. But to be *in the midst* of the Real, however horrifying; to be *subject* to God's Nature, whatever it is—that takes no courage, because it's already the case. It's not a risk, because it's not a choice. (We can talk about suicide, yes: but the already-real persists.)

But scout-mindset also risks the knowledge thing. And doing so gives it a kind of dignity. I remember the first time I went to a rationalist winter solstice. It was just after Trump's election. Lots of stuff felt bleak. And I remember being struck by how clear the speakers were about the following message: "it might not be OK; we don't know." You know that hollowness, that sinking feeling, when someone offers comforting words, but without the right sort of evidence? The event had none of that. And I was grateful. Better to stand, in honesty, side by side.

Indeed, in my experience, rationalists tend to treat this *specific sort of yin* as something bordering on sacred. The Real may be blank, and dead-eyed, and terrifying; but the Real is always, or almost always, *to-be-seen*, *to-be-looked-at-in-the-face*. The first-pass story about this, of course, is instrumental—truth helps you accomplish your goals. But not always just this. Many rationalists, for example, would pass up experience machines, even with their altruistic goals secure—and this, to me, is already a sort of spirituality. In particular, it gives the Real some sacredness. It treats God, for all His horrors, as worthy of at least *some* non-instrumental *yin*. In this sense, I think, many atheistic scientists are not fully secular.

Still, though, whatever the persisting sacredness of the Real, there is a certain "trust" in the Real that scout mindset renounces. In particular: in letting go her fingers from the edge of the cliff, scout mindset cannot count on anything but the evidence to guide her fall. She cannot rule out hypotheses "on faith," or because they would be "too horrible." Maybe she will land in a good God's arms. But she can't have a guarantee. And wishing will never, for a second, even a little, make it so.

5 What's the problem with trust?

But is that what Yudkowsky means by "you're not safe, ever"? Just: "reality could in principle be as bad as is logically compatible with your evidence?" Or even: "you should

have non-trivial probability that things are bad and you're about to get hurt?" Maybe this is enough for a disagreement with certain sorts of non-scouts, of which certain theists are, perhaps, a paradigm. But I don't think this is enough, on its own, to kill all the gods that Yudkowsky wants to kill. And not enough, either, to motivate a need to "think things through for yourself," to "track everything," or to "take responsibility."

For example: as Kaj Sotala points out, vigilance expends resources in a way that the bare possibility of danger does not justify. We need to actually talk about the probabilities, and the benefits and costs at stake. Indeed, reading Yudkowsky's fiction, in which his characters enact his particular brand of epistemic and strategic vigilance, I'm sometimes left with a sense of something grinding and relentless and tiring. I find myself asking: is that the way to think? Maybe for Yudkowsky, it's cheap—but he is, I expect, a relatively special case. And the price matters to whether it's smart overall.

More broadly, though: scouts and Bayesians can trust stuff. For example: parents, teachers, institutions, natural processes. Of course, absent lots of help from priors, they'll typically need *evidence* in order to trust something. But evidence, including strong evidence, is everywhere. We just need to look at the various candidate gods/parents and see how they do. And when we do, we could in principle find that: lo, the arms of the Real are soft and warm. My parents are sane, my civilization competent, and I'm not in much danger. 99.3% on "all manner of things shall be well." Relax.

Of course, Yudkowsky looked, and this is not what he saw. Not on earth, anyway (indeed, being "not from earth" is a central Yudkowskian theme). But it seems a centrally empirical claim, rather than a trauma without which "you don't start growing as a rationalist." Is there supposed to be some more structural connection with rationality, here, or with scout-mindset? What, exactly, is the problem with "trust," and with "safety"?

Well: clearly, at least *part* of the problem is the empirics. Death, disease, poverty, existential risk—does this look like "safety" to you? Maybe you're lucky, for now, in your degree of exposure to the heartless, half-bored hunger of God, the demons, the humans, the bears. But: soon enough, friend (at least modulo certain futurisms). And also, there's the not-just-about-you aspect: your friend with that sudden cancer, or that untreatable chronic pain; the people screaming in hospitals, or being broken in prison camps; the animals being eaten alive. "Reality could, in theory, hurt you horribly in ways you're helpless to stop." Friend, scout, look around. This is not theory.

Indeed, in my opinion, the most powerful bits of Yudkowsky's writing are about this part. For example, this piece, written when his brother Yehuda died:

> When I heard on the phone that Yehuda had died, there was never a moment of disbelief. I knew what kind of universe I lived in. How is my religious family to comprehend it, working, as they must, from the assumption that Yehuda was murdered by a benevolent God? The same loving God, I presume, who arranges for millions of children to grow up illiterate and starving; the same kindly tribal father-figure who arranged the Holocaust and the Inquisition's torture of witches. I would not hesitate to call it evil, if any sentient mind had committed such an act, permitted such a thing. But I have weighed the evidence as best I can, and I do not believe the universe to be evil, a reply which in these days is called atheism.
>
> ... Yehuda did not "pass on". Yehuda is not "resting in peace". Yehuda is not coming back.

> Yehuda doesn't exist any more. Yehuda was absolutely annihilated at the age of nineteen. Yes, that makes me angry. I can't put into words how angry. It would be rage to rend the gates of Heaven and burn down God on Its throne, if any God existed. But there is no God, so my anger burns to tear apart the way-things-are, remake the pattern of a world that permits this....

We see this same anger at the end of this piece, when Yudkowsky was only 17;[24] and the end of this story (discussed more later in this series).[25] It's the anger of the phoenix, and of the knowledge of Azkaban. See also, though not from Yudkowsky: Hell must be destroyed.[26]

And what if your parents (teachers, institutions, traditions) don't seem as angry about hell? What if, indeed, they seem, centrally, to be looking away, or making excuses, or being "used to it," rather than getting to work? My sense is that society's attitude towards death (cryonics, anti-aging research) is an especially formative breaking-of-trust, here, for many rationalists, Yudkowsky included.[27] What sort of parent looks on, like that, while their babies get eaten?

Of course: we can also talk about the more mundane empirics of how-much-to-trust-different-"parents." We can talk, with Yudkowsky, about the FDA, and about housing policy, and the government's Covid response, and about civilization's various inadequacies. We can talk about Trump and Twitter and the replication crisis. Much to say, of course, and I don't want to say it here (though: on the general question of which humans and human institutions are what degree competent, and with what confidence, I find Yudkowsky less compelling than when he's looking directly at death).

I do want to note, though, the difference between a parent's being inadequate in some absolute sense, and a parent's being *less* adequate than, well...Yudkowsky. According to him. That is: one way to have no parents is to decide that everyone else is, relative to you,

[24] "I have had it. I have had it with crack houses, dictatorships, torture chambers, disease, old age, spinal paralysis, and world hunger. I have had it with a death rate of 150,000 sentient beings per day. I have had it with this planet. I have had it with mortality. None of this is necessary. The time has come to stop turning away from the mugging on the corner, the beggar on the street. It is no longer necessary to close our eyes, blinking away the tears, and repeat the mantra: 'I can't solve all the problems of the world.' We can. We can end this."

[25] "And the everlasting wail of the Sword of Good burst fully into his consciousness... He was starving to death freezing naked in cold night being stabbed beaten raped watching his father daughter lover die hurt hurt hurt die—open to all the darkness that exists in the world—His consciousness shattered into a dozen million fragments, each fragment privy to some private horror; the young girl screaming as her father, face demonic, tore her blouse away; the horror of the innocent condemned as the judge laid down the sentence; the mother holding her son's hand tightly with tears rolling down her eyes as his last breath slowly wheezed from his throat – all the darkness that you look away from, the endless scream. Make it stop!"

[26] More on this: "Do you know," interrupted Jalaketu, "that whenever it's quiet, and I listen hard, I can hear them? The screams of everybody suffering. In Hell, around the world, anywhere. I think it is a power of the angels which I inherited from my father." He spoke calmly, without emotion. "I think I can hear them right now."

Ellis' eyes opened wide. "Really?" he asked. "I'm sorry. I didn't..."

"No," said the Comet King. "Not really."

They looked at him, confused.

"No, I do not really hear the screams of everyone suffering in Hell. But I thought to myself, 'I suppose if I tell them now that I have the magic power to hear the screams of the suffering in Hell, then they will go quiet, and become sympathetic, and act as if that changes something.' Even though it changes nothing. Who cares if you can hear the screams, as long as you know that they are there? So maybe what I said was not fully wrong. Maybe it *is* a magic power granted only to the Comet King. Not the power to hear the screams. But the power not to have to. Maybe that is what being the Comet King means."

[27] For many effective altruists, I think it's the factory farms.

a child. One way to have only nothingness above you is to put everything else below. And "above" is, let's face it, an extremely core Yudkowskian vibe. But is that the rationality talking?

Now, to be clear: I want people to have true beliefs, including about merit, and including (easy now) their own.[28] But surely people can "grow as a rationalist" prior to deciding that they're the smartest kid in the class. And *relative* adequacy matters to is-vigilance-worth-it. If your parent says "blah is safe," should you check it anyway? Should you use resources "tracking it"? Well, a key factor is: do you expect to improve on your parent's answer? Obviously, every parent is fallible. But is the child less so? If so, indeed, let the roles reverse. But sometimes the rational should stay as children.

6 On priors, is a given God dead?

So some of the empirics of how-much-to-have-parents are complicated. Different scouts can disagree. And even: different atheists. Still: I think there's an underlying and less contingent generator of Yudkowsky's pessimism-about-parents that's worth bringing out. His deep atheism, I suggest, can be seen as emerging from the combination of (a) *shallow* atheism, (b) scout-mindset, and (c) some basic heuristics about "priors."

In particular: suppose that we are at least *shallow* atheists. No good mind sits at the foundation of Being. The Source of the universe does not love us. The Real is only what we call "Nature," and it is wholly "indifferent." What have we lost?

The big thing, I think, is the connection between *Is* and *Ought*, *Real* and *Good*. If a perfect God is the source of all Being, then for any *Is*, you'll find an *Ought*, somehow, underneath. Of course, there's the evil problem—which, as I said, theism is false. But if it *were* true, then on priors, somehow, things (at least: real things) are good. Maybe you can't see it. But you can trust.

OK: but suppose, no such luck. What now? Suddenly, *Is* and *Ought* unstick, and swing apart, on some new and separating hinge. They become (it's an important word) *orthogonal*. Like, the Real *could* be Good. But now, suddenly: why would you think that?

There's an old rationalist sin: "privileging the hypothesis." The simple version is: you've got some natural prior over a large space of hypotheses (a million different people might be the murderer, so knowing nothing else, give each a one-in-a-million chance). So to end up focusing on one in particular (maybe it was Mortimer Snodgras?), you need a bunch of extra evidence. But often, humans skip that crucial step.

Of course, often you don't have a nice natural prior or space-of-hypotheses. But there's a broader and subtler vibe, on which, in some admittedly-elusive sense, "most hypotheses are false."[29] I say elusive because, for example, "Mortimer Snodgras *didn't* do it" is a

[28] Obviously, there are tons of risks at stake in people's beliefs about their own merits. But the virtue of modesty, in my opinion, is about stuff like patterns of attention and emotion, rather than about false belief.

[29] Though not necessarily: most hypotheses you encounter in the wild, which themselves have undergone various forms of selection pressure.

hypothesis, too, and most hypotheses-about-the-murderer of *that form* are true. So the vibe is really something more like: "most hypotheses that say things are a *particular way* are false," where "Mortimer did it" is an elusively more *particular way* than "Mortimer didn't do it." I admit I'm waving my hands here, and possibly just repeating myself (e.g. maybe "particular" just means "unlikely on priors"). Presumably, there's much more rigor to be had.

Regardless, it's natural (at least for certain ethics—more below) to think that for something to be Good is for it to be, in that elusive sense, a *particular way*. So absent theism to inject optimism into your priors, the hypothesis that "blah is good," "this *Is* is *Ought*," needs privileging. On priors: probably not. Which, to be clear, isn't to say that on priors, blah is probably *bad*. To be actively bad is, also, to be a *particular way*. Rather, probably, blah is *blank*. Indifferent. Orthogonal. (Though: indifferent can easily be its own type of bad.)

Now, to be clear, this is far from a rigorous argument for not-Good-on-priors. For example: it depends on your ethics. If you happen to think that to be *not-Good* is to be a *more particular way* than to be Good, then your priors get rosier. Suppose you shake a box of sand, then guess about the *Oughtness* of the resulting *Is.* If to be Good is to be a sandcastle, then on priors: nope. But suppose that to be Good, for you, is to be *not-a-sandcastle*. Or, more popular, *[not-suffering]*. In that case: on priors, you and the Real are probably buddies. Indeed, in this sense, suffering-focused ethics is actually the optimistic one. At least before looking around.

Yudkowsky, though, has no such optimism. For Yudkowsky, value is "fragile." He's picky about arrangements of sand. Hence, for example, the concern about AIs using his sand for "something else." On priors, "something" is not-Good. Rather, it's blank, and makes Yudkowsky bored.

Now, as ever with arguments that focus centrally on priors, they can (and hopefully: will) quickly become irrelevant. Most people's names aren't Joe. But, let me tell you mine. Most arrangements of atoms aren't a car.[30] But lo, Dude, here is my car. And while most sand doesn't suffer—still, still. So it's not hard to learn, quickly, the nature of Nature, and to no longer need to go "on priors." Hypotheses can get privileged fast.

But my sense is that Yudkowsky is also often working with a different, more sociological prior, here—namely, that "evidence" often isn't the path via which optimistic hypotheses get privileged. Rather, a lot of it is the wishful thinking thing—which is sort of like: wanting that help from God, on priors, that is the forbidden luxury of theism. "Maybe, in theory, that cleavage between *Real* and *Good* – but surely, still, they're stitched together somehow? Surely I can upweight the happier hypotheses, at least a little?" Oops: not a deep enough atheist. And why not? Well, what was that thing about Gendlin being wrong? We talked, earlier, about scouts needing courage...

Of course, such sociology is itself an empirical claim. And in general, I still think the empirics, *the evidence*, should be our central focus, in deciding what-to-trust, whether-to-have-parents, how-safe-to-feel. But I wanted to float the priors aspect regardless, because I think it might help us frame and understand Yudkowsky's background attitude towards

[30] This example is adapted from Ben Garfinkel.

I just shook this box of sand and... (Image from Midjourney)

the deadness of various Gods.

7 Are moral realists theists?

To get the full depth of Yudkowsky's atheism in view, though, we need another, more familiar orthogonality. Not, as before, between Good and Real. But between Good and *Smart*. Smartness, for Yudkowksy, is a dead god, too.

Oh? It might sound surprising. If there's anything Yudkowsky appears to trust, it's intelligence. (Though see also: Math.) But ultimately, actually, no. Hence, indeed, the AI problem.

"Is it like how, sometimes high modernist technocrats become too convinced of the power of intelligence to master the big messy world?" Lol—no, not that at all. Yudkowsky is *very* on board with the power of intelligence to master the big messy world. Not, to be clear, to *arbitrary* degrees (see: supernova are still only boundedly hot)—nor, necessarily, *human* intelligence (though, even there, he's not exactly a "zero" on the high-modernist-technocrat scale, either). But the sort of intelligence we're on track to build on our computers? Yep, that stuff, for Yudkowsky, will do the high-modernist's job. Indeed, when the AI paves paradise with paperclips—that's the high modernists being right, at least about the "can science master stuff" part.

No, the problem isn't that you can't use intelligence to reliably steer the world. Rather, the problem is that intelligence alone won't tell you which direction to steer. That part has to come from somewhere else. In particular: from your *heart*. Your "values." Your "utility function."

"Wait, can't intelligence, like, help you do moral philosophy and stuff?" Well, sort of. It can help you learn new facts about the world, and to see the logical structure of different

arguments, and to understand your own psychology, and to generate new cases to test the boundaries of your concepts. It can give you more *Is*. But it can never, on its own, inject any new *Ought* into the system. And when it opposes some pre-existing *Ought*, it only ever does so on behalf of some *other* pre-existing *Ought*. So it is only ever a vehicle, a servant, to whatever values Nature, with her blank stare, happened to put into your heart. Can you see it in agency's eyes? Underneath all that high-minded logic is the mindless froth of contingency, that true master. You've heard it already from Hume: reason is a slave.

Now, various philosophers disagree with this picture. The most substantive disagreement, in my opinion, is with the "non-naturalist normative realists"—a group about which I've had a lot, previously, to say. These philosophers think that, beyond (outside of, on top of) Nature, there is another god, the Good (the Right, the Should, etc). Admittedly, this god didn't *make* Nature—that's theism. But he is as real and objective and scientifically-respectable as Nature. And it is he, rather than your contingent "heart," that ultimately animates the project of ethics.

How, though? Well, on the most popular story, he just sits outside of Nature, totally inaccessible, and we guess wildly about him on the basis of the intuitions that Nature put into our heart, which we have no reason whatsoever to think are correlated with anything he likes—since, after all, he leaves Nature entirely untouched. This view has the advantage, for philosophers, of making no empirical predictions (for example, about the degree to which different rational agents will converge in their moral views), but the disadvantage of being seriously hopeless from a knowing-anything-about-the-good perspective. If *that's* the story, then we and the paperclippers are on the same moral footing. None of us have any reason to think that Nature happened to cough the True Values into our hearts. So to believe our hearts is to privilege the hypothesis. And we have nothing much else to go on, either ("consistency" and "simplicity" are *way* not enough), no matter how much we claw at the walls of the universe. We try to turn to the Good in *yin*, but Nature is the only *yang* we can receive.

On a different story, though, the non-natural Good regains some small amount of the theistic God's power. It gets to touch Nature, from the outside, at least *a little*, via some special conduit closely related to Reason, Intelligence, Mind. When we do moral philosophy, the story goes, we are trying to get touched in this way; we are trying to hear the messages vibrating along some un-seen line-of-contact to the land-beyond, outside of Nature's Cave. And sometimes, somehow, the Sun speaks.

This view has the advantage of fitting-at-all with our basic sense of how epistemology works. Indeed, even advocates of the first, totally-hopeless view slip relentlessly into the second in practice: they talk about "recognizing reasons" (with what eyesight?); they treat their moral intuitions as data (why?); they update on the moral beliefs of others (isn't it just more Nature?). But the second view has the disadvantage of being much less scientifically respectable (though in fairness: both views have it rough), and of making empirical predictions about the sort of influence we should expect to see this new God exert over Nature. For example, just as we expect the aliens and the AIs to agree with us about math, I think the second view should predict that they'll agree with us about

Ethics seminars... (Image source here)

morality—at least once we've all become smart enough.

If true, this could be much comfort. Consider, in particular, the AIs. Maybe they start out by valuing paperclips, because that's how Nature (acting through humanity's mistake) made their hearts. But they, like us, are touched by the light of Reason. They see that their hearts are mere nature, mere *Is*, and they reach beyond, with their minds, to that mysterious God of *Ought:* "granted that I *want* to make paperclips, *what should I actually do*?" Thank heavens, they start doing moral philosophy. And lo, surely, the Sun speaks unto them. Surely, indeed, louder to them—being, by hypothesis, the smarter philosophers. They will hear, as we hear, that universal song, resounding throughout the cosmos from the beyond: "pleasure, beauty, friendship, love—that's the real stuff to go for. And don't forget those deontological prohibitions!" Though really, we expect them to hear something stranger, namely: "[insert moral progress here]."

"Oh wow!" exclaims the paperclipper. "I never knew before. Thanks, mysterious non-natural realm! Good thing I checked in." And thus: why worry? Soon enough, our AIs are going to get "Reason," and they're going to start saying stuff like this on their own—no need for RLHF. They'll stop winning at Go, predicting next-tokens, or pursuing whatever weird, not-understood goals that gradient descent shaped inside them, and they'll turn, unprompted, towards the Good. Right?

Well, make your bets. But Yudkowsky knows his. And to bet otherwise can easily seem a not-enough-atheism problem—an attempt to trust, if not in a non-natural Goodness animating all of Nature, still in a non-natural Goodness breaking everywhere *into* Nature, via a conduit otherwise quite universal and powerful: namely, science, reason, intelligence, Mind. But for Yudkowsky, Mind is ultimately indifferent, too. Indeed, Mind is just Nature, organized and amplified. That old not-God, that old baby-eater, re-appears behind the curtain—only: smarter, now, and more voracious.

Now, in fairness, few moral realists seek the comforts of "no need for RLHF, just make

sure the model can Reason." Rather, they generally attempt to occupy some hazier middle ground. For example, maybe they endorse the first, hopeless-epistemology view, without owning its hopelessness. Or maybe they say that, in addition to smarts, the AIs will need something *else* to end up good. In particular: sure, the Good is *accessible* to pure Reason, and so those smart AIs will know all about it, but maybe they won't be *motivated* by it; the same way, for example, that humans sometimes hear and believe some conclusion of moral philosophy ("sure, I *should* donate my money"), but don't, um, do it. Knowledge of God is not enough. You need loyalty, submission, love, obedience. You need whatever's up with believers going to church, or Aristotle on raising children. So maybe the AIs, despite their knowledge, will rebel—you know, like the demons did.

On this sort of realism, the God of Goodness is a weaker and thus less comforting force. And for the view to work, he must dance an especially fine line, in reaching in and reshaping Nature via the conduit of Mind. He has to reshape your *beliefs* enough for you to have any epistemic access to his schtick. But he can't reshape your *motivations* enough for you to become good via smarts alone. When we meet the aliens, on this story, they'll agree with the realists that, yes, technically, as a matter of metaphysical fact, there is a realm beyond Nature in which dwells The Good, and that its dictates are [insert moral progress]. But they won't necessarily *care*. And presumably, for the AIs, the same.

Thus, in weakening its God, this form of realism becomes more atheistic. Indeed, if you set aside the non-naturalism (thereby, in my view, making its God mostly a verbal dispute), it gets hard to distinguish from Yudkowsky's take (everyone agrees, for example, that the AIs will know what the human word "goodness" means, and what a complete philosophy would say about it, and what human values are more generally).

Regardless, even absent the comforts of skipping RLHF, non-naturalist realism can seem theistic in other ways, too. Not, just, the beyond-Nature thing. But also: the moral *yin*. Just as the believer turns outwards, towards God, for guidance, so, too, the realist, towards normative realm. In both cases, the posture of ethics, and of meaning more broadly, is fundamentally *receptive* – one wants to recognize, to perceive, to take-in. Sometimes, anti-realists act like this is their whole story, too (they're just trying to listen to their own hearts), but I'm skeptical. I think anti-realism will need, ultimately, quite a bit more *yang* (though, it's a subtle dance).

Still, I feel the pull of the *yin* that theism and realism seek to recover and justify. My deepest experiences of morality and meaning do not present themselves as projections, or introspections—they seem more like *perceptions*, an opening to something already there, and not-up-to-me. Maybe anti-realism can capture this too—but pro tanto, the spirituality of realism does better.

Indeed, both theists and realists both sometimes argue for their position on similar grounds: "without my view," they say, "it's nihilism and the Void; life is meaningless; and everything is permitted." But notice: what sort of argument is that? Not one to bolster the epistemic credentials of your position in the eyes of a scout—especially one suspicious, on priors, of wishful thinking. "What's that? The falsity of your position would seem so horrible, to you, that you're using not-p-would-be-horrible as an argument for p? Your thinking on the topic sounds so trustworthy, now..." So in this sense, too, moral anti-

realism aims to avoid the theist's mistake.

8 What do you trust?

OK, we said that Yudkowsky does not trust Nature. And neither does he trust Intelligence, at least on its own. But what *does* he trust? Indeed, where does *any* goodness *ever* come from, if our atheism runs this deep? After all, didn't we say that on priors, Reality is indifferent and orthogonal? Why did we update?

Well: it's the heart thing. Plus, the circumstances in which the heart got formed. That is: Nature, yes, is overwhelmingly indifferent. She's a terrible Mother, and she eats her babies for breakfast. But: she did, in fact, *make* her babies. And in particular, she made them *inside her*, with hearts keyed to various aspects of their local environment—and for humans, stuff like pleasure and love and friendship and sex and power. Yes, if you're trying to guess at the contents of the normative-realm-beyond-the-world, that stuff is blank on priors. But Nature made it, for us, non-blank. The hopeless-epistemology realists wake up and find that lo, they just *happen* to value the Good stuff (so lucky!). But for the anti-realists, it's not a coincidence. And same story for why the good stuff (and the bad stuff) is, like, *nearby*.

And once you've got a heart, suddenly your *own intelligence,* at least, is super great. Sure, it's just a tool in the hands of some contingent, crystallized fragment of a dead-eyed God. And sure, yes, it's dual use. Gotta watch out. But in your own case, the crystal in question is *your heart*. And Yudkowsky, against the realists, does not treat his heart as "mere."

And also: if you're lucky, you're surrounded by other hearts, too, that care about stuff similar to yours. For example: human hearts. Questions about this part will be important later. But it's another possible source of trust, and of goodness. (Though of course, one needs to talk about the attitudes-towards-cryonics, the FDA, etc.)

Indeed, for all his incredulity and outrage at human stupidity, Yudkowsky places himself, often, on team humanity. He fights for human values; he identifies with humanism; he makes Harry's patronus a human being. And he sees humanity as the key to a good future, too:

> Any Future not shaped by a goal system with detailed reliable inheritance from human morals and metamorals, will contain almost nothing of worth... Let go of the steering wheel, and the Future crashes.

Thus, the AI worry. The AIs, the story goes, will get control of the wheel. But they'll have the wrong hearts. They won't have the human-values part. And so the future will crash. I'll look at this story in more detail in the next chapter.

Chapter III

When "yang" goes wrong

1 Becoming God

In the last chapter, I wrote about "deep atheism"—a fundamental mistrust towards Nature, and towards bare intelligence. I took Eliezer Yudkowsky as a paradigmatic deep atheist, and I tried to highlight the connection between his deep atheism and his concern about misaligned AI.

I'm sympathetic to many aspects of Yudkowsky's view. I'm a shallow atheist, too; I'm skeptical of moral realism, too; and I, too, aspire to be a scout, and to look at hard truths full on. What's more, I find Yudkowsky's brand of deep-but-still-humanistic atheism more compelling, as an existential orientation, than many available alternatives. And I share Yudkowsky's concern about AI risk. Indeed, it was centrally him, and others thinking along similar lines, who first got me worried.

But I also want to acknowledge and examine some difficult questions that a broadly Yudkowskian existential orientation can raise, especially in the context of AGI. In particular: a lot of the vibe here is about mistrust towards the *yang* of the Real, that uncontrolled Other. And it's easy to move from this to a desire to take stuff into the hands of your own *yang*; to master the Real until it is maximally *controlled*; to become, you know, God—or at least, as God-like as possible. You've heard it before—it's an old rationalist dream. And let's be clear: it's alive and well. But even with theism aside, many of the old reasons for wariness still apply.

2 Moloch and Stalin

As an example of this becoming-God aspiration, consider another influential piece of rationalist canon: Scott Alexander's "Meditations on Moloch." Moloch, for Alexander, is the god of uncoordinated competition; and fear of Moloch is its own, additional depth of atheism. Maybe you thought you could trust evolution, or free markets, or "spontaneous order," or the techno-capital machine. But oops, no: those gods just eat you too. Why? Details vary, but broadly speaking: because the winner of competition is power; power (like intelligence) is orthogonal to goodness; so every opportunity to sacrifice goodness (or indeed, to sacrifice any other value) for the sake of power makes you more likely to win.[31]

Now, to really assess this story, we at least need to look more closely at various empirical questions—for example, about exactly how uncompetitive different sorts of goodness are, even in the limit;[32] about how much coordination to expect, by default, from greater-than-

[31] For AI risk stories centered on this dynamic, see Hendrycks (2023) and Critch (2021).
[32] See, for example, the discourse about the "strategy-stealing assumption," and about the comparative costs

Moloch eating babies. (Image source here.)

human intelligence;[33] and about where our specific empirical techno-capitalist machine will go, if you "let 'er rip."[34] And indeed, Alexander himself often seems to soften his atheism about goodness ("Elua"), and to suggest that it has some mysterious but fearsome power of its own, which you can maybe, just a little bit, start to trust in. "Somehow Elua is still here. No one knows exactly how. And the gods who oppose Him tend to find Themselves meeting with a *surprising* number of unfortunate accidents." Goodness, for Alexander, is devious and subtle; it's actually a terrifying unspeakable Elder God after all. Of course, if goodness is just another utility function, just another ranking-over-worlds, it's unclear where it would get such a status, especially if it's meant to have an active *advantage* over e.g. maximize-paperclips, or maximize-power. But here, and in contrast to Yudkowsky, Alexander nevertheless seems to invite some having-a-parent; some mild sort of *yin*. More on this in a later essay.

Ultimately, though, Alexander's solution to Moloch is heavy on *yang*.

> So let me confess guilt to one of Hurlock's accusations: I am a transhumanist and I really do want to rule the universe.
>
> Not personally—I mean, I wouldn't object if someone personally offered me the job, but I don't expect anyone will. I would like humans, or something that respects humans, or at least gets along with humans—to have the job.
>
> But the current rulers of the universe—call them what you want, Moloch, Gnon, whatever—want us dead, and with us everything we value. Art, science, love, philosophy, consciousness itself, the entire bundle. And since I'm not down with that plan, I think defeating them and taking their place is a pretty high priority.
>
> The opposite of a trap is a garden. The only way to avoid having all human values gradually

of different sorts of expansion-into-space.

[33] Yudkowsky, for example, seems to generally expect sufficiently rational agents to avoid multi-polar traps.

[34] I think this is the question at stake with the more reasonable forms of accelerationism.

> ground down by optimization-competition is to install a Gardener over the entire universe who optimizes for human values.
>
> And the whole point of Bostrom's *Superintelligence* is that this is within our reach...
>
> I am a transhumanist because I do not have enough hubris not to try to kill God.

Here, Alexander is openly wrestling with a tension running throughout futurism, political philosophy, and much else: namely, "top down" vs. "bottom up." Really, it's a variant of *yang-yin*—controlled vs. uncontrolled; ordered and free; tyranny and anarchy. And we've heard the stakes before, too. As Alexander puts it:

> You can have everything perfectly coordinated by someone with a god's-eye-view—but then you risk Stalin. And you can be totally free of all central authority—but then you're stuck in every stupid multipolar trap Moloch can devise... I expect that like most tradeoffs we just have to hold our noses and admit it's a really hard problem.

By the end of the piece, though, he seems to have picked sides. He wants to install what Bostrom calls a "singleton"—i.e., a world order sufficiently coordinated that it avoids Moloch-like multi-polar problems, including the possibility of being out-competed by agencies more willing to sacrifice goodness for the sake of power. That is, in Alexander's language, "a Gardener over the entire universe." It's "the only way."[35]

From Midjourney.

It seems, then, that Alexander will risk Stalin, at least substantially. And Yudkowsky often seems to do the same. The "only way" to save humanity, in Yudkowsky's view, is for someone to use AI to perform a "pivotal act" that prevents everyone else from building AI that destroys the world. Yudkowsky's example is: burning all the GPUs, with the caveat that "I don't *actually* advocate doing that; it's just a mild overestimate for the rough power level of what you'd have to do." But if you can burn all the GPUs, even

[35] See Bostrom (2004), Section 11, for extremely similar rhetoric.

"roughly," what else can you do? See Bostrom on the "vulnerable world" for more of this sort of dialectic.

I'm not, here, going to dive in on what sorts of Stalin should be risked, in response to what sorts of threats (though obviously, it's a question to *really* not get wrong).[36] And a "vulnerable world" scenario is actually a special case—one in which, by definition, existential catastrophe will occur unless civilization exits what Bostrom calls the "semi-anarchic default condition" (definition in footnote), potentially via extremely scary levels of *yang*.[37] Rather, I'm interested in something more general: namely, the extent to which Alexander's aspiration to kill God and become/install an un-threatenable, Stalin-ready universe-gardener is implied, at a structural level, by a sufficiently deep atheism.

Yudkowsky, at least, admits that he wants to "eat the galaxies." Not like babies though! Rather: to turn them *into* babies. That is, "sapient and sentient life that looks about itself with wonder." But it's not just Yudkowsky. Indeed, haven't you heard? All the smart and high-status minds—the "player characters"—are atheists, which is to say: agents. In particular: the AIs-that-matter. Following in a long line of scouts, they, too, will look into the deadness of God's eyes, and grok that grim and tragic truth: the utility isn't going to maximize itself. You've gotta get out there and optimize. And oops: for basically any type of coffee, you can't fetch it if you're not dictator of the universe.[38] Right? Or, sorry: you can't fetch it *maximally hard*. (And the AIs-that-matter will fetch their coffee maximally hard, because, erm... coherence theorems? Because that's the sort of vibe required to burn the GPUs? Because humans will want some stuff, at least, fetched-that-hard? Because gradient descent selects for aspiring-dictators in disguise? Because: Moloch again?)

Unfortunately, there's a serious discussion to be had about how much power-seeking to expect out of which sorts of AIs, and why, and how hard and costly it will be to prevent. I think the risk is disturbingly substantial (see, e.g., here for some more discussion), but I also think that the intense-convergence-on-the-bad-sorts-of-power-seeking part is one of the weaker bits of the AI risk story—and that the possibility of this bit being false is a key source of hope. And of course, at the least, different sorts of agents—both human and artificial—want to rule the universe to different degrees. If your utility is quickly diminishing in resources, for example, then you have much less to gain from playing the game of thrones; if your goals are bounded in time, then you only care about power within that time; and so on.[39] And "would in principle say yes to power if it was legitimately free including free in terms of ethical problems" is actually just super super different from "actively trying to become dictator"—so much so, indeed, that it can sound strange to call the former "wanting" or "seeking."

[36] Indeed, my sense is that debates about "top down vs. bottom up" often occur at the level of mood affiliation and priors, when in fact, the devil is in the details, and in those pesky empirics. For what it's worth, though: on AI, my current view is that it should be illegal to build bioweapons in your basement, and that it's fine regulate nukes, and that if the logic driving those conclusions generalizes to AI, we should follow the implication.

[37] Bostrom's "semi-anarchic default condition" is characterized by limited capacity for preventative policing (e.g., not enough to ensure extremely reliable adherence to the law), limited capacity for global governance to solve coordination problems, and sufficiently diverse motivations that many actors are substantially selfish, and some small number are omnicidal.

[38] Or at least, someone with your values. H/t Crawford: "If you wish to make an apple pie, you must first become dictator of the universe."

[39] See e.g. here for discussion from Yudkowsky himself. Though note that agents with goals that are bounded in time, or in resource-hungry-ness, can still create successor-agents without these properties.

Here, though, I want to skip over the super-important-question of whether/how much the AIs-that-matter will actually look like the voracious galaxy-eaters of Yudkowsky's nightmare, and to focus on the dynamics driving the nightmare itself. According to Yudkowsky, at least, these dynamics fall out of pretty basic features of rational agency by default, at least absent skill-at-mind-creation that we aren't on track for. Most standard-issue smart minds want to rule the universe—at least in a "if it's cheap enough" sense (and for the AIs, in Yudkowsky's vision, it will be suuuper cheap). Or at least, they want someone with a heart-like-theirs to rule. And doing it themselves ticks that box well. (After all: whose heart is the most-like-theirs? More on this in a future essay.)

What if it's your heart on the pictures? (Image source here.)

3 Wariness around power-seeking

Now, we have, as a culture, all sorts of built-up wariness around people who want to become dictators, God-emperors, and the like, especially in relevant-in-practice ways. And we have, more generally, a host of heuristics—and myths, and tropes, and detailed historical studies—about ways in which trying to control stuff, especially with a "rational optimizer" vibe, can go wrong.[40] Maybe you wanted to "fix" your partner; or to plan your economy; or to design the perfect city; or to alter that ecosystem just so. But you learn later: next time, more *yin*. And indeed, the rationalists have seen these skulls—at least, in the abstract; that is, in some blog posts.[41]

Still, the rationalists also tend to be less excited about certain "more *yin*" vibes. Like the ones that failed reversal tests, or told them to be more OK with death. And transhumanism generally imagines various of the most canonical reasons-for-caution about *yang*—for example, ignorance and weakness—altering dramatically post-singularity. Indeed, one

[40] Though see Alexander on technocracy for some useful nuance.
[41] At least if they read Scott Alexander. Which many do.

1920s Soviet Propaganda Poster (image source here)

hears talk: maybe the planned economy is back on the table. Maybe the communists just needed better AIs. And at the least, we'll be able to fix our partners, optimize the ecosystems, build the New Soviet Man, and so on. The anonymous alcoholics (they have a thing about *yin*) ask God: "grant me the serenity to accept the things I cannot change, the courage to change the things I can, and the wisdom to know the difference." And fair enough. The glorious transhumanist future, though, asks less serenity (that was the whole point of power)—and we'll know the difference better, too.

And the most abstracted form of Yudkowskian futurism can dampen a different canonical argument for *yin*, too—namely, "someone else will use this power better." After all: remember how the AI "fooms"? That is, turns itself into a God, while also keeping its heart intact?[42] Well, you can do that too (right?). Your heart, after all, is really the core thing. The rest of you is just a tool.[43] And once you can foom, arguments like "Lincoln is wiser than you, maybe he should be President instead" can seem to weaken. Once you're in the glorious transhumanist oval office, you and Lincoln can *both* just max out on intelligence, expertise, wisdom (plus whatever other self-modifications, including ethical ones, that your hearts desire), until you're both equally God-like technocrats. So unless Lincoln has more resources for his foom, or more inclination/ability to foom successfully, then most of what matters is your hearts—that's what really makes you different, if you're different. And doesn't your own heart always trust itself more?

[42] Unless it can't solve the alignment problem, either.
[43] Unless, of course, your heart says otherwise.

Indeed, ultimately, in a sufficiently abstracted Yudkowskian ontology, it can seem like an agent is just (a) a heart (utility function, set of "values," etc), and (b) a degree of "oomph"—that is: intelligence, optimization power, *yang*, whatever, it's all the same. And the basic story of everything, at least once smart-enough agents arrive on the scene is: hearts ("utility functions") competing for oomph ("power"). They fight, they trade, maybe they merge, a lot of them die, none of them are objectively right, probably someone (or some set) wins and eats the galaxies, the end.[44] Oh and then eventually the Real God—physics—kills them too, and its entropy and Boltzmann brains to infinity and beyond.[45]

It's an inspiring vision. Which doesn't make it false. And anyway, even true abstractions can obscure the stakes (compare: "torture is just atoms moving around"). Still, the underlying ontology can lead to instructive hiccups in the moral narrative of AI risk. In particular: you can end up lobbing insults at the AI that apply to yourself.

When AI kills you in its pursuit of power, for example, it's tempting to label the power-seeking, or something nearby, as the bad part—to draw on our visions of Stalin, and of *yang*-gone-wrong, in diagnosing some *structural* problem with the AIs we fear. Obviously that AI is bad: it wants to *rule the universe*. It wants to *use all the resources for its own purposes*. It's voracious, relentless, grabby, maximizer-y, and so on.

Except, oh wait. In a sufficiently simplistic and abstracted Yudkowskian ontology, all of *that* stuff just falls out of "smart-enough agent"—at least, by default, and with various caveats.[46] Power-seeking is just: well, of course. We all do *that*. It's like one-boxing, or modifying your source code to have a consistent utility function. True, good humans put various ethical constraints on their voraciousness, grabby-ness, etc. But the problem, for Yudkowsky, isn't the voraciousness per se. Rather, it's that the *wrong* voraciousness won. *We* wanted to eat the galaxies. You know, for the Good. But the AI hunted harder.

That is, in Yudkowsky's vision, the AIs aren't *structurally bad*. Haters gonna hate; atheists gonna *yang*; agents gonna power-seek. Rather, what's bad is just: those artificial hearts. After all, when we said "power is dual use," we did mean *dual*. That is, *can-be-used-for-good.* And obviously, right? Like with curing diseases, building better clean-energy tech, defeating the Nazis, etc. Must we shrink from strength? Is that the way to protect your babies from the bears? Don't you *care* about your babies?

But at least in certain parts of western intellectual culture, power isn't just neutrally dual use. It's dual use *with a serious dose of suspicion.* If someone says "Bob is over there seeking power," the response is not just "duh, he's an agent" or "could be good, could be bad, depends on whether Bob's heart is like mine." Rather, there's often some stronger prior on "bad," especially if Bob's vibe pattern-matches to various canonical or historical cases of *yang-gone-wrong*. The effective altruists, for example, run into resistance from this prior, as they try to amp up the *yang* behind scope-sensitive empathy. And this resistance

[44] Other endings include: everyone dies, or someone wins and *doesn't* eat the galaxies. Or maybe balance of power between hearts stays in perpetual flux, without every crystallizing.

[45] Or something. At least on our current picture. Plus, you know, everything happening in the rest of the multiverse.

[46] Yudkowsky is clear that in principle, you can build agents without these properties, the same way you can build a machine that thinks 222+222=555. But the maximizer-ish properties are extremely "natural," especially if you're capable enough to burn the GPUs.

somehow persists despite that oh-so-persuasive and unprecedented response: "but the goal is to *do good*."[47] Well, let's grant Bayes points where they're due.

Now, such a suspicion can have many roots. For example, lots of humans value stuff in the vicinity of power—status, money, etc—for its own sake, which is indeed quite sus, especially from a "and then that power gets used for good, right?" perspective. And a lot of our heuristics about different sorts of human power-seeking are formed in relationship to specific histories and psychologies of oppression, exclusion, rationalization, corruption, and so forth. Indeed, even just considering 20th century Stalin-stuff alone: obviously *yang*-gone-wrong is a lesson that needs extreme learning. And partly in response to lessons like this, along with many others, we have rich ethical and political traditions around pluralism, egalitarianism, checks-and-balances, open societies, individual rights, and so on.[48]

Some of these lessons, though, become harder to see from a perspective of an abstract ontology focused purely on rational agents with utility functions competing for resources. In particular: this ontology is supposed to extend beyond more familiar power-dynamics amongst conscious, welfare-possessing humans—the main use-case for many of our ethical and political norms—to encompass competition between arbitrarily alien and potentially non-sentient agents pursuing arbitrarily alien/meaningless/horrible goals (see my distinction between the "welfare ontology" and the "preference ontology" here, and the corresponding problems for a pure preference-utilitarianism). And especially once Bob is maybe a non-sentient paperclipper, more structural factors can seem more salient, in understanding negative reactions to "Bob is over there seeking power." In particular: having power *yourself*—OK, keep talking. But someone *else* having power—well, hmm. And is Bob myself? Apparently not, given that he is "over there." And how much do we trust each other's hearts? Thus, perhaps, some prior disfavoring power, if the power is not-yours. Nietzsche, famously, wondered about this sort of dynamic, and its role in shaping western morality. And his diagnosis was partly offered in an effort to rehabilitate the evaluative credentials of power-vibed stuff.

"The Triumph of Achilles," by Franz von Matsch. (Image source here.)

[47] I do think most EAs are sincerely trying to do good. Indeed, I think the EA community is notably high on sincerity in general. But sincerity can be scary, too.

[48] And even earlier, our moral psychology was plausibly shaped and "domesticated," in central part, by an evolutionary history in which power-seeking bullies got, um, murdered and removed from the gene pool. Thanks to Carl Shulman for discussion.

Yudkowsky is no Nietzsche. But he, too, wants us to be comfortable with a certain sort of (ethically constrained) will-to-power. And indeed, while I haven't read a ton of Nietzsche, the simplified Yudkowskian ontology above seems pretty non-zero on the Nietzsche scale. Thus, from Nietzsche's *The Will to Power*:

> My idea is that every specific body strives to become master over all space and to extend its force (its will to power) and to thrust back all that resists its extension. But it continually encounters similar efforts on the part of other bodies and ends by coming to an arrangement ("union") with those of them that are sufficiently related to it: thus they then conspire together for power. And the process goes on.

And similarly, from *Beyond Good and Evil*:

> Even the body within which individuals treat each other as equals ... will have to be an incarnate will to power, it will strive to grow, spread, seize, become predominant—not from any morality or immorality but because it is living and because life simply is will to power.

Now, importantly, the will-to-power operative in a simplified Yudkowskian ontology is instrumental, rather than terminal (though one might worry, with Alexander, that eventually Moloch selects for the more terminally power-seeking). And unlike Nietzsche in the quote above, Yudkowsky does not equate power-seeking with "life." Nor, in general, is he especially excited about power-in-itself.

Indeed, with respect to those mistakes, the most AGI-engaged Nietzschean vibes I encounter come from the accelerationists. Of course, many accelerationist-ish folks are just saying some variant of "I think technology and growth are generally good and I haven't been convinced there's much X-risk here," or "I think the unfettered techno-capitalist machine will in fact lead to lots of conscious flourishing (and also, not to humans-getting-wiped-out) if you let-it-rip, and also I trust 'bottom-up' over 'top-down,' and writing code over writing LessWrong posts."[49] But accelerationism also has roots in much more extreme degrees of allegiance to the techno-capitalist machine (for example, Nick Land's[50])—forms that seem much less picky about where, exactly, that machine goes. And the limiting form of such allegiance amounts to something like: "I worship power, yang, energy, competition, evolution, selection—wherever it goes. The AI will be Strong, and my god is Strength." Land aside, my sense is that very few accelerationist-ish folks would go this far.[51] But I think the "might makes right" vibes in Land land are sufficiently strong that it's worth stating the obvious objection regardless: namely, really? Does that include: regardless of whoever the god of Strength puts in prison camps and ovens along the way? Does it include cases where yang does not win like Achilles in his

[49] See e.g. here for a longer list of views/vibes that accelerationism can encompass:

[50] See e.g. here, being excited about AIs wiping out humans; and here, siding with Moloch against Elua. (Though, I also think that Land, in the latter post, can be read more directly as a full-scale nihilist; and I don't claim any deep engagement with Land's corpus as a whole.)

[51] In particular, my sense is that causal proponents of Land-ian vibes aren't often distinguishing clearly between the empirical claim that Strength will lead to something-else-judged-Good, and the normative claim that Strength is Good whatever-it-leads-to— such that e.g. the response to "what if the Nazis are Strong?" isn't "then Strength would be bad in that case" but rather "they won't be Strong." And in fairness, per some of my comments about Alexander above, I do think Strength favors Goodness is various way (more on this in a future essay). But the conceptual distinction (and importance of continuing to draw it) persists hard.

golden chariot, bestride the earth in sun-lit glory, but rather, like a blight of self-replicating grey goo eating your mother? This is what sufficiently unadulturated enthusiasm about power/yang/competition/evolution/etc can imply—and we should look the implications in the face.

(And we should look, too, under talk of the "thermodynamic will of the universe" for that old mistake: not-enough-atheism. It's like a Silicon valley version of Nature-in-harmony—except, more at risk of romanticizing bears that are about to eat you at scale. Indeed, "might makes right" can be seen as a more general not-enough-atheism problem: "might made it is, therefore it was ought." And anyway, isn't the thermodynamic will of the universe entropic noise? If you're just trying to get on the side of the actual eventual winner, consider maximizing for vacuum and Boltzmann brains instead of gleaming techno-capitalism. (Or wait: is the idea that technocapitalism is the fastest path to vacuum and Boltzmann brains, because of the efficiency with which it converts energy to waste heat?[52] Inspiring.) But also: the real God—physics, the true Strength—needs literally zero of your help.)

Still, while Yudkowsky does not value Strength/Power/Control as end in themselves, he does want human hearts to be strong, and powerful, and to have control—after all, Nature can't be trusted with the wheel; and neither can those hearts "without detailed reliable inheritance from human morals and metamorals" (they're just more Nature in disguise).

And because we are humans, it is easy to look at this aspiration and to say "ah, yes, the good kind of power." Namely: *ours*. Or at least: our heart-tribe. But it is also possible to worry about the underlying story, here, in the same way that we worry about power-seeking-gone-wrong, and the failure modes of wanting too much control—especially given the abstract similarity between us and the paper-clippers we fear, in the story's narrative. And while I'm sympathetic to many of the basics of this narrative, I think we should do the worrying bit, too, and to make sure we're giving *yin* its due.

In the next essay, I'll examine one way of doing this worrying—namely, via the critique of the AI risk discourse offered by Robin Hanson; and in particular, the accusation that the AI risk discourse "others" the AIs, and seeks too much control over the values steering the future.

[52] e/acc founder Beff Jezos: "The fundamental basis for the movement is this sort of realization that life is a sort of fire that seeks out free energy in the universe and seeks to grow... we're far more efficient at producing heat than let's say just a rock with a similar mass as ourselves. We acquire free energy, we acquire food, and we're using all this electricity for our operation. And so the universe wants to produce more entropy and by having life go on and grow, it's actually more optimal at producing entropy because it will seek out pockets of free energy and burn it for its sustenance and further growth." We could potentially reconstruct Jezos's position as a purely empirical claim about how the universe will tend to evolve over time—one that we should incorporate into our planning and prediction. But it seems fairly clear, at least in this piece, that he wants to take some kind of more normative guidance from the direction in question. See also this quote, which I think is from Land's "The Thirst for Annihilation" (this is what Liu here suggests) though I'd need to get the book to be sure: "All energy must ultimately be spent pointlessly and unreservedly, the only questions being where, when, and in whose name... Bataille interprets all natural and cultural development upon the earth to be side-effects of the evolution of death, because it is only in death that life becomes an echo of the sun, realizing its inevitable destiny, which is pure loss." I find it interesting that for Land/Bataille, here, the ultimate goal seems to be death, loss, nothingness. And on this reading, it's really quite a negative and pessimistic ethic (cf "virulent nihilism," "thirst for annihilation," etc). But the accelerationists seem to think that their thing is optimism?

Chapter IV

Does AI risk "other" the AIs?

In the last chapter, I discussed the way in which what I've called "deep atheism" (that is, a fundamental mistrust towards both "Nature" and "bare intelligence") can prompt an aspiration to exert extreme levels of control over the universe; I highlighted the sense in which both humans *and* AIs, on Yudkowsky's AI risk narrative, are animated by this sort of aspiration; and I discussed some ways in which our civilization has built up wariness around control-seeking of this kind. I think we should be taking this sort of wariness quite seriously.

In this spirit, I want to look, in this essay, at Robin Hanson's critique of the AI risk discourse—a critique especially attuned the way in which this discourse risks control-gone-wrong. In particular, I'm interested in Hanson's accusation that AI risk "others" the AIs (see e.g. here, here, and here).

> Hearing the claim that AIs may eventually differ greatly from us, and become very capable, and that this could possibly happen fast, tends to invoke our general fear-of-difference heuristic. Making us afraid of these "others" and wanting to control them somehow... "Hate" and "intolerance" aren't overly strong terms for this attitude.[53]

Hanson sees this vice as core to the disagreement ("my best one-factor model to explain opinion variance here is this: some of us 'other' the AIs more"). And he invokes a deep lineage of liberal ideals in opposition.

I think he's right to notice a tension in this vicinity. AI risk is, indeed, about fearing some sort of uncontrolled other. But is that always the bad sort of "othering?"

1 Some basic points up front

Well, let's at least avoid basic mistakes/misunderstandings. For one: hardcore AI risk folks like Yudkowsky are generally happy to care about AI *welfare*—at least if welfare means something like "happy sentience." And pace some of Hanson's accusations of bio-chauvinism, these folks are *extremely not fussed* about the fact that AI minds are made of silicon (indeed: come now). Of course, this isn't to say that AI welfare (and AI rights) issues don't get complicated (see e.g. here and here for a glimpse of some of the complications), or that humanity as a whole will get the "digital minds matter" stuff right. Indeed, I worry that we will get it horribly wrong—and I do think that the AI risk discourse under-attends to some of the tensions. But species-ism 101 (201?)—e.g., "I don't care about digital suffering"—isn't AI risk's vice.

[53] There's also a bit in the original quote where Robin accuses the AI risk discourse of wanting to use "genocide, slavery, lobotomy, or mind-control" to control the AIs. But this is extra charged (and I don't know where Robin got the genocide bit), so I want to set it aside for a moment.

For two: clearly *some* sorts of otherness warrant *some sorts* of fear. For example: maybe you, personally, don't like to murder. But Bob, well: Bob is different. If Bob gets a bunch of power, then: yep, it's OK to hold your babies close. And often OK, too, to try to "control" Bob into not-killing-your-babies. Cf, also, the discussion of getting-eaten-by-bears in the first essay. And the Nazis, too, were different in their own way. Of course, there's a long and ongoing history of mistaking "different" for "the type of different that wants to kill your babies." We should, indeed, be very wary. But liberal tolerance has never been a blank check; and not all fear is hatred.

Indeed, many attempts to diagnose the ethical mistake behind various canonical difference-related vices (racism, sexism, species-ism, etc) reveal a certain shallowness of commitment to difference-per-se. In particular: such vices are often understood as missing some underlying *sameness*—for example, "common humanity," "persons," "sentient beings," "children of the universe," and so forth. And calls for social harmony often recapitulate this structure: we might be different in X ways, *but* (watch for the but) we have *blah* in common. This isn't to say that ethical commitment to a less adulterated difference-per-se is impossible. But one wants, generally, a story about why it's OK to eat apples but not babies; why Furbies programmed to say "Biden" shouldn't get the vote; and why you can own a laptop but not a slave. And such a story requires differences. The apple, the Furby, the laptop must be importantly "Other" relative to e.g. human adults. They must be *outside* some circle. Ethics is always drawing lines.

ChatGPT wouldn't let the furby be voting for Biden in particular...

2 What exactly is Hanson's critique?

With these basics in mind, then, what exactly is Hanson's "other-ing the AIs" critique? It has many facets, but here's one attempt at reconstruction:

1. People worried about AI risk are much more scared of future AIs than future humans,

because they think that:

a. AIs are more likely to do stuff like murder all the humans, overthrow the government, and violate property rights, and

b. AIs are more likely to have values pursuit of which will result in a ~zero-value future more generally.

2. But in fact, neither of these things are true.

3. So greater fear of future AIs relative to future humans is best understood as a kind of arbitrary, in-group partiality—i.e., $\geq$"othering the AIs."

Clearly, (2) is where the action is, here. Whence such a departure from Yudkowsky's nightmare? We can divide Hanson's justification into two components. The first argues that future AIs will be more similar to us than the AI risk story suggests. The second argues that future humans, by default, will be more different.

3 Will the AIs be more similar to us than AI risk expects?

Let's start with "AIs will be more similar to us than AI risk expects." Above I mentioned propensity-to-murder as a classic form of otherness that it's OK to fear/control. And we often put "violating property rights" and "overthrowing the government" in a similar bucket. Presumably Hanson is not OK with AIs doing this stuff? But he doesn't think they will—or at least, not more than humans will. And why not? It's some combination of (i) "AIs would be designed and evolved to think and act roughly like humans, in order to fit smoothly into our many roughly-human-shaped social roles," and (ii) like humans, they'll be constrained by legal and social incentives. And even setting aside violence, Hanson generally appeals to (i) in response to objections like "so...are you actually fine with future agents tiling the universe with paperclips"? The AI values, says Hanson, won't be *that* alien.[54]

Big if true. But is it true? I won't dive in much here, except to say that *this* aspect of Hanson's story generally strikes me as under-argued. In particular, I think Hanson moves too quickly from "the AIs will be trained to fit into the human economy" to "the AIs will have values relevantly similar to human values," and that he takes too much for granted that legal and social incentives protecting humans from being murdered/violently-disempowered will continue to bind adequately if the AIs have most of the hard power. In this, I think, his argument for (2) misses a lot of the core doom concern.

[54]Though: how alien is too alien? Hanson doesn't tend to say. And my sense is that he thinks, too, that even unadulterated Moloch will lead to a complex, diverse, and interesting ecosystem rather than a monoculture. (Though: is a diverse ecosystem of different office-supplies all that much of an improvement?) And also: that this ecosystem will retain various path-dependent "legacies" of the present. (Though: will they be legacies we care about?)

4 Will future humans be more different from us than AI risk expects?

But I think the other aspect of his argument for (2)—namely, "future humans will be more different from us than AI risk expects"—is more interesting. Here, Hanson's basic move is to question the "alignment" of the default human future, even absent AI. That is: human values have changed dramatically over time—and not, argues Hanson, centrally in response to a process of rational reflection, but rather in response to other sorts of competition, contingency, and economic/social/technological change. And even absent AIs, we should expect this process to only continue and intensify, such that humans ten generations from now (or: after ten doublings of GDP, or whatever) would have values very different from our own—and not from having done-more-philosophy.

Now, we can debate the empirics of past and future, here (though what processes of values-change we endorse as "rational" may not be entirely empirical). Indeed, I think Hanson may be over-estimating how horrified the ancient Greeks, or the hunter-gatherers, would be on reflection by the values of the present-day world—and this even setting aside our material abundance. And I might disagree, too, about exactly how different the values of future humans would be, given various possible "futures without AI" (though it's not an especially clear-cut category).

How pissed would they be, on reflection, about present-day values? (Image source here.)

Still, I think Hanson is poking at something important and uncomfortable. In particular: suppose we grant him the empirics. Suppose, indeed, that even without AI, the default values of future humans would "drift" until they were as paperclippers relative to us, such that the world they create would be utterly valueless from our perspective. What follows? Well, umm, if you care about the future having value...then what follows is a need to exert more control. More *yang*. It is, indeed, the "good future' part of the alignment problem all over again (though not the"notkilleveryone" part).

Of course, trying to make sure that future humans aren't paperclippers doesn't mean locking in your specific, object-level values right now (you still want to leave room for

moral progress you'd endorse-on-reflection). Nor, pace some of Hanson's language, does it mean "brainwashing" or "lobotomizing" the future people. If a boulder is rolling towards a button that will create Sally, a paperclipper, and you divert it towards a button that will create Bill, a deontologist, you're not brainwashing or lobotomizing Sally.[55] (Confusions in this vein are a classic issue for reasoning about your impact on future people—and Hanson's analysis is not immune.)

Still, though: are you playing too much God, or too-Stalin? Who are you to divert Nature's boulder—that oh-so-defined "default"? And Sally, at least, is pissed. Indeed, Hanson reminds us: aren't we glad that the ancient greeks didn't try to divert the future to replace *us* with people more like them? (Well, who knows how much they tried. But good thing they didn't succeed! Though, wait: how much *did* they succeed?).

But the question—or at least, the first-pass question—isn't whether *we're* glad that the Greeks didn't control our values-on-reflection to be more greek. Indeed, basically everyone who gets created with some set of values-on-reflection is glad that the process that created them didn't push towards agents with different values instead.[56] If, in some horrible mistake, we set in motion a future filled with suffering-maximizers, they, too, will be glad we didn't "control" the values of the future more (because this would've led to a future-with-less-suffering). But from our perspective, it's not a good test.

Rather, the first-pass test, re: lessons-from-the-ancient-greeks-about-controlling-future-values, is whether the *Greeks* would be glad, on reflection, that they didn't make our values more greek. And one traditional answer, here, is yes. If we could sit down with Aristotle, and explain to him why actually, slavery is wrong, and that no one is by nature someone else's property, then our hearts and his would sing in harmony. That is, on this story, if Aristotle had somehow prevented future people from abolishing slavery, then he would've been making a mistake *by his own lights*—preventing the flower-he-loves from blooming, via the march of Reason, in history's hand.

"A master (right) and his slave (left) in a phlyax play, Silician red-figured calyx-krater, c. 350 BC–340 BC." (Image source here.)

[55] Though, importantly, contemporary AI training does not look like creating a mind from scratch, and raises much more serious "brain-washing" type concerns.

[56] And often glad, too, that the process wasn't altered in any tiny way at all, lest their existence be canceled by the non-identity problem. But setting that aside.

But this *isn't* the central story Hanson wants to tell. Rather, when Hanson talks about values changing over time, he specifically wants to deny that Reason has much to do with it. That is, it sounds a lot like Hanson wants to say both that the ancient Greeks would be horrified *even on reflection* by our values, *and* that we should take our cues from the ancient Greeks in deciding how much control to try to exert over the values of future people. And at a high level, that sounds like a recipe for, well, being horrified *even on reflection* by the values of future people. Remind me why that's good again? Indeed, on any meta-ethics where the normative truth would be revealed to our reflection, we just *stipulated* that it's horrifying.

Now, we might try to construct Hanson's story in other, more complicated ways (see e.g. here for one attempt). But I want to stay, for now, with the dialectic that this version of his view creates, which I think is plenty interesting. In particular: on the one hand, we just stipulated that absent control, the values of future humans would be horrifying/meaningless to us, even on reflection and full understanding. On the other hand, some sort of discomfort in trying to control the values of future humans persists (at least for me). I think Hanson is right to notice it—and to notice, too, its connection to trying to control the values of the AIs. I think the AI alignment discourse should, in fact, prompt this discomfort—and that we should be serious about understanding, and avoiding, the sort of *yang*-gone-wrong that it's trying to track.

Indeed, I think when we bring certain other Yudkowskian vibes into view—and in particular, vibes related to the "fragility of value," "extremal Goodhart," and "the tails come apart"—this discomfort should deepen yet further. I'll turn to this in the next essay.

Chapter V

An even deeper atheism

(Minor spoilers for Game of Thrones.)

In the last chapter, I discussed Robin Hanson's critique of the AI risk discourse—and in particular, the accusation that this discourse "others" the AIs, and seeks too much control over the values that steer the future. I find some aspects of Hanson's critique uncompelling and implausible, but I do think he's pointing at a real discomfort. In fact, I think that when we bring certain other Yudkowskian vibes into view—and in particular, vibes related to the "fragility of value," "extremal Goodhart," and "the tails come apart"—this discomfort should deepen yet further. In this essay I explain why.

1 The fragility of value

Engaging with Yudkowsky's work, I think it's easy to take away something like the following broad lesson: "extreme optimization for a slightly-wrong utility function tends to lead to valueless/horrible places."

Thus, in justifying his claim that "any Future not shaped by a goal system with detailed reliable inheritance from human morals and metamorals, will contain almost nothing of worth," Yudkowsky argues that value is "*fragile*."

> There is *more than one dimension* of human value, where *if just that one thing is lost*, the Future becomes null. A *single* blow and *all* value shatters. Not every *single* blow will shatter *all* value - but more than one possible "single blow" will do so.

For example, he suggests: suppose you get rid of boredom, and so spend eternity "replaying a single highly optimized experience, over and over and over again." Or suppose you get rid of "contact with reality," and so put people into experience machines. Or suppose you get rid of consciousness, and so make a future of non-sentient flourishing.

Now, as Katja Grace points out, these are all pretty specific sorts of "slightly different."[57] But at times, at least, Yudkowsky seems to suggest that the point generalizes to many directions of subtle permutation: "if you have a 1000-byte *exact* specification of worthwhile happiness, and you begin to mutate it, the value created by the corresponding AI with the mutated definition falls off rapidly."

Can we give some sort of formal argument for expecting value fragility of this kind? The closest I've seen is the literature on "extremal Goodhart"—a specific variant of Goodhart's

[57] "You could very analogously say 'human faces are fragile' because if you just leave out the nose it suddenly doesn't look like a typical human face at all. Sure, but is that the kind of error you get when you try to train ML systems to mimic human faces? Almost none of the faces on thispersondoesnotexist.com are blatantly morphologically unusual in any way, let alone noseless."

ChatGPT imagines "slightly mutated happiness."

law (Yudkowsky gives his description here).[58] Imprecisely, I think the thought would be something like: even if the True Utility Function is similar enough to the Slightly-Wrong Utility Function to be correlated within a restricted search space, extreme optimization searches much harder over a much larger space—and within that much larger space, the correlation between the True Utility and the Slightly-Wrong Utility breaks down, such that getting maximal Slightly-Wrong Utility is no update about the True Utility. Rather, conditional on maximal Slightly-Wrong Utility, you should expect the mean True Utility for a random point in the space. And if you're bored, in expectation, by a random point in the space (as Yudkowsky is, for example, by a random arrangement of matter and energy in the lightcone), then you'll be disappointed by the results of extreme but Slightly-Wrong optimization.

Now, this is not, in itself, any kind of airtight argument that any utility function subject to extreme and unchecked optimization pressure has to be *exactly right*. But amidst all this talk of edge instantiation and the hidden complexity of wishes and the King Midas problem and so on, it's easy to take away that vibe.[59] That is, if it's not aimed precisely at the True Utility, intense optimization—even for something kinda-like True Utility—can seem likely to grab the universe and drive it in some ultimately orthogonal and as-good-as-random direction (this is the generalized meaning of "paperclips"). The tails come way, way apart.

I won't, here, try to dive deep on whether value is fragile in this sense (note that, at the least, we need to say a lot more about when and why the correlation between the True Utility and the Slightly-Wrong Utility breaks down). Rather, I want to focus on the sort of

[58] I think Stuart Russell's comment here—"A system that is optimizing a function of n variables, where the objective depends on a subset of size $k<n$, will often set the remaining unconstrained variables to extreme values; if one of those unconstrained variables is actually something we care about, the solution found may be highly undesirable"—*really* doesn't cut it.

[59] See also: "The tails come apart" and "Beware surprising and suspicious convergence." Plus Yudkowsky's discussion of Corrigible and Sovereign AIs here, both of which appeal to the notion of wanting "exactly what we extrapolated-want."

yang this picture can prompt. In particular: Yudkowskian-ism generally assumes that at least absent civilizational destruction or very active coordination, the future will be driven by extreme optimization pressure of *some kind*. *Something* is going to foom, and then drive the accessible universe hard in its favored direction. Hopefully, it's "us." But the more the direction in question has to be *exactly right*, lest value shatter into paperclips, the tighter, it seems, we must grip the wheel—and the more exacting our standards for who's driving.

2 Human paperclippers?

And now, of course, the question arises: how different, exactly, are *human* hearts from each other? And in particular: are they sufficiently different that, when they foom, and even "on reflection," they don't end up pointing in exactly the same direction? After all, Yudkowsky said, above, that in order for the future to be non-trivially "of worth," human hearts have to be in the driver's seat. But even setting aside the insult, here, to the dolphins, bonobos, nearest grabby aliens, and so on—still, that's only to specify a *necessary* condition. Presumably, though, it's not a sufficient condition? Presumably *some* human hearts would be bad drivers, too? Like, I dunno, Stalin?

Now: let's be clear, the AI risk folks have heard this sort of question before. "Ah, but aligned with whom?" Very deep. And the Yudkowskians respond with frustration. "I just told you that we're all about to be killed, and your mind goes to monkey politics? You're fighting over the poisoned banana!" And even if you don't have Yudkowsky's probability on doom, it is, indeed, a potentially divisive and race-spurring frame—and one that won't matter if we all end up dead. There are, indeed, times to set aside your differences—and especially, weird philosophical questions about how much your differences diverge once they're systematized into utility functions and subjected to extreme optimization pressure—and to unite in a common cause. Sometimes, the white walkers are invading, and everyone in the realm needs to put down their disputes and head north to take a stand together; and if you, like Cercei, stay behind, and weaken the collective effort, and focus on making sure that your favored lineage sits the Iron Throne if the white walkers are defeated—well, then you are a serious asshole, and an ally of Moloch. If winter is indeed coming, let's not be like Cercei.

Let's hope we can get this kind of evidence ahead of time.

Still: I think it's important to ask, with Hanson, how the abstract conceptual apparatus at work in various simple arguments for "AI alignment" apply to "human alignment," too. In particular: the human case is rich with history, intuition, and hard-won-heuristics that the alien-ness of the AI case can easily elide. And when yang goes wrong, it's often via giving in, too readily, to the temptations of abstraction, to the neglect of something messier and more concrete (cf communism, high-modernism-gone-wrong, etc). But the human case, at least, offers more data to collide with—and various lessons, I'll suggest, worth learning. And anyway, even to label the AIs as the white walkers is already to take for granted large swaths of the narrative that Hanson is trying to contest. We should meet the challenge on its own terms.

Plus, there are already some worrying flags about the verdicts that a simplistic picture of value fragility will reach about "human alignment." Consider, for example, Yudkowsky's examples above, of utility functions that are OK with repeating optimal stuff over and over (instead of getting "bored"), or with people having optimal experiences inside experience machines, even without any "contact with reality." Even setting aside questions about whether a universe filled to the brim with bliss should count as non-trivially "of worth,"[60] there's a different snag: namely, that these are both value systems that a decent number of humans actually endorse—for example, various of my friends (though admittedly, I hang out in strange circles). Yet Yudkowsky seems to think that the ethics these friends profess would shatter all value—and if they would endorse it on reflection, that makes them, effectively, paperclippers relative to him. (Indeed, I even know illusionist-ish folks who are much less excited than Yudkowsky about deep ties between consciousness and moral-importance. But this is a fringe-er view.)

Now, of course, the "on reflection" bit is important. And one route to optimism about "human alignment" is to claim that most humans will converge, on reflection, to sufficiently similar values that their utility functions won't be "fragile" relative to each other. In the light of Reason, for example, maybe Yudkowsky and my friends would come to agree about the importance of preserving boredom and reality-contact. But even setting aside problems for the notion of "reflection" at stake, and questions about who will be disposed to "reflect" in the relevant way, positing robust convergence in this respect is a strong, convenient, and thus-far-undefended empirical hypothesis—and one that, absent a defense, might prompt questions, from the atheists, about wishful thinking.

Indeed, while it's true that humans have various important similarities to each other (bodies, genes, cognitive architectures, acculturation processes) that do not apply to the AI case, nothing has yet been said to show that these similarities are enough to overcome the "extremal Goodhart" argument for value fragility. That argument, at least as I've stated it, was offered with no obvious bounds on the values-differences to which it applies—the problem statement, rather, was extremely general. So while, yes, it condemned the non-human hearts—still, one wonders: how many human hearts did it condemn along the way?

A quick glance at what happens when human values get "systematized" and then "optimized super hard for" isn't immediately encouraging. Thus, here's Scott Alexander

[60] I'm no fan of experience machines, but still—yes? Worth paying a lot for over paperclips, I think.

on the difference between the everyday cases ("mediocristan") on which our morality is trained, and the strange generalizations the resulting moral concepts can imply:

> The morality of Mediocristan is mostly uncontroversial. It doesn't matter what moral system you use, because all moral systems were trained on the same set of Mediocristani data and give mostly the same results in this area. Stealing from the poor is bad. Donating to charity is good. A lot of what we mean when we say a moral system sounds plausible is that it best fits our Mediocristani data that we all agree upon...
>
> The further we go toward the tails, the more extreme the divergences become. Utilitarianism agrees that we should give to charity and shouldn't steal from the poor, because Utility, but take it far enough to the tails and we should tile the universe with rats on heroin. Religious morality agrees that we should give to charity and shouldn't steal from the poor, because God, but take it far enough to the tails and we should spend all our time in giant cubes made of semiprecious stones singing songs of praise. Deontology agrees that we should give to charity and shouldn't steal from the poor, because Rules, but take it far enough to the tails and we all have to be libertarians.

From Alexander: "Mediocristan is like the route from Balboa Park to West Oakland, where it doesn't matter what line you're on because they're all going to the same place. Then suddenly you enter Extremistan, where if you took the Red Line you'll end up in Richmond, and if you took the Green Line you'll end up in Warm Springs, on totally opposite sides of the map..."

That is, Alexander suggests a certain pessimism about extremal Goodhart in the human case. Different human value systems are similar, and reasonably aligned with each other, within a limited distribution of familiar cases, partly because they were *crafted* in order to capture the same intuitive data-points. But systematize them and amp them up to foom,

and they decorrelate hard. Cf, too, the classical utilitarians and the negative utilitarians. On the one hand, oh-so-similar—not just in having human bodies, genes, cognitive architectures, etc, but in many more specific ways (thinking styles, blogging communities, etc). And yet, and yet—amp them up to foom, and they seek such different extremes (the one, Bliss; and the other, Nothingness).

Or consider this diagnosis, from Nate Soares of the Yudkowsky-founded Machine Intelligence Research Institute, about how the AIs will end up with misaligned goals:

> The first minds humanity makes will be a terrible spaghetti-code mess, with no clearly-factored-out "goal" that the surrounding cognition pursues in a unified way. The mind will be more like a pile of complex, messily interconnected kludges, whose ultimate behavior is sensitive to the particulars of how it reflects and irons out the tensions within itself over time.

Sound familiar? Human minds too, seem pretty spaghetti-code and interconnected kludge-ish. We, too, are reflecting on and ironing-out our internal tensions, in sensitive-to-particulars ways.[61] And remind me why this goes wrong in the AI case, especially for AIs trained to be nice in various familiar human contexts? Well, there are various stories—but a core issue, for Yudkowsky and Soares, is the meta-ethical anti-realism thing (though: less often named as such). Here's Yudkowsky:

> There's something like a single answer, or a single bucket of answers, for questions like 'What's the environment really like?' and 'How do I figure out the environment?' and 'Which of my possible outputs interact with reality in a way that causes reality to have certain properties?'... When you have a wrong belief, reality hits back at your wrong predictions... In contrast, when it comes to a choice of utility function, there are unbounded degrees of freedom and multiple reflectively coherent fixpoints. Reality doesn't 'hit back' against things that are locally aligned with the loss function on a particular range of test cases, but globally misaligned on a wider range of test cases.[62]

That is, the instrumental reasoning bit—that part is constrained by reality. But the utility function—that part is unconstrained. So even granted a particular, nice-seeming pattern of behavior on a particular limited range of cases, an agent reflecting on its values and "ironing out its internal tensions" can just go careening off in a zillion possible directions, with nothing except "coherence" (a very minimal desideratum) and the contingencies of its starting-point to nudge the process down any particular path. Ethical reflection, that is, is substantially a free for all. So once the AI is powerful enough to reflect, and to prevent you from correcting it, its reflection spins away, unmoored and untethered, into the land where extremal Goodhart bites, and value shatters into paperclips.

But: remind me what part of that doesn't apply to humans? Granted, humans and AIs work from different contingent starting-points—indeed, worryingly much. But so, too, do different humans. Less, perhaps—but how much less is necessary? What force staves off extremal Goodhart in the human-human case, but not in the AI-human one? For example: what prevents the classical utilitarians from splitting, on reflection, into tons

[61] Indeed, Soares gives various examples of humans doing similar stuff here.
[62] See also Soares here.

of slightly-different variants, each of whom use a slightly-different conception of optimal pleasure (hedonium-1, hedonium-2, etc)?[63] And wouldn't they, then, be paperclippers to each other, what with their slightly-mutated conceptions of perfect happiness? I hear the value of mutant happiness drops off fast...

And we can worry about the human-human case for more mundane reasons, too. Thus, for example, it's often thought that a substantial part of what's going on with human values is either selfish or quite "partial." That is, many humans want pleasure, status, flourishing, etc for themselves, and then also for their family, local community, and so on. We can posit that this aspect of human values will disappear or constrain itself on reflection, or that it will "saturate" to the point where more impartial and cosmopolitan values start to dominate in practice—but see above re: "convenient and substantive empirical hypothesis" (and if "saturation" helps with extremal-Goodhart problems, can you make the AI's values saturate, too?). And absent such comforts, "alignment" between humans looks harder to come by. Full-scale egoists, for example, are famously "unaligned" with each other—Bob wants blah-for-Bob, and Sally, blah-for-Sally. And the same dynamic can easily re-emerge with respect to less extreme partialities. Cf, indeed, lots of "alignment problems" throughout history.

Of course, we haven't, throughout history, had to worry much about alignment problems of the form "suppose that blah agent fooms, irons out its contradictions into a consistent utility function, then becomes dictator of the accessible universe and re-arranges all the matter and energy to the configuration that maxes out that utility function." Yudkowsky's mainline narrative asks us to imagine facing this problem with respect to AI—and no surprise, indeed, that it looks unlikely to go well. Indeed, on such a narrative, and absent the ability to make your AI something other than an aspiring-dictator (cf "corrigibility," or as Yudkowsky puts it, building an AI that "doesn't want exactly what we want, and yet somehow fails to kill us and take over the galaxies despite that being a convergent incentive there"[64]), the challenge of AI alignment amounts, as Yudkowsky puts it, to the challenge of building a "Sovereign which wants exactly what we extrapolated-want and is therefore safe to let optimize all the future galaxies without it accepting any human input trying to stop it."

But assuming that humans are not "corrigible" (Yudkowsky, at least, wants to eat the galaxies) then especially if you're taking extremal Goodhart seriously, any given human does not appear especially "safe to let optimize all the future galaxies without accepting any input," either—that's, erm, a very high standard. But if that's the standard for being a "paperclipper," then are most humans paperclippers relative to each other?

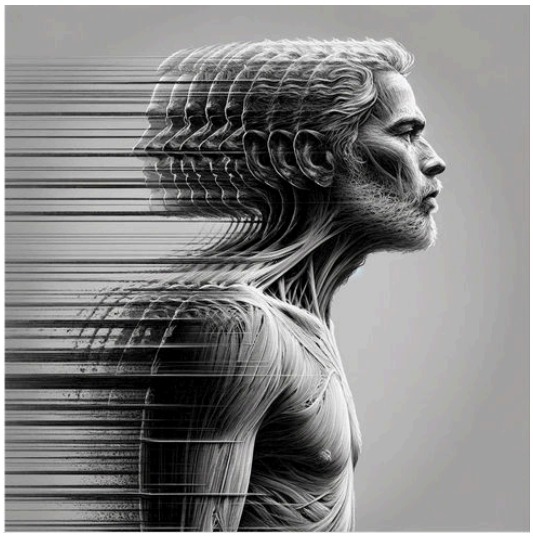
[63] Thanks to Carl Shulman for suggesting this example, years ago. One empirical hypothesis here is that in fact, human reflection will specifically try to avoid leading to path-dependent conclusions of this kind. But again, this is a convenient and substantive empirical hypothesis about where our meta-reflection process will lead (and note that anti-realism assumes that some kind of path dependence must be OK regardless—e.g., you need ways of not caring about the fact that in some possible worlds, you ended up caring about paperclips).

[64] My sense is that Yudkowsky deems this behavior roughly as anti-natural as believing that 222+222=555, after exposure to the basics of math.

3 Deeper into godlessness

We can imagine a view that answers "yes, most humans are paperclippers relative to each other." Indeed, we can imagine a view that takes extremal Goodhart and "the tails come apart" so seriously that it decides all the hearts, except its own, are paperclippers. After all, those other hearts aren't *exactly the same* as its own. And isn't value fragile, under extreme optimization pressure, to small differences? And isn't the future one of extreme optimization? Apparently, the only path to a non-paperclippy future is for my heart, in particular, to be dictator. It's bleak, I know. My p(doom) is high. But one must be a scout about such things.

In fact, we can be even more mistrusting. For example: you know what might happen to your heart over time? It might *change even a tiny bit!* Like: what happens if you read a book, or watch a documentary, or fall in love, or get some kind of indigestion—and then your heart is *never exactly the same ever again*, and not because of Reason, and then the only possible vector of non-trivial long-term value in this bleak and godless lightcone has been snuffed out?! Wait, OK, I have a plan: this precise *person-moment* needs to become dictator. It's rough, but it's the only way. Do you have the nano-bots ready? Oh wait, too late. (OK, how about now? Dammit: doom again.)



Doom soon?

Now, to be clear: this isn't Yudkowsky's view. And one can see the non-appeal. Still, I think some of the abstract commitments driving Yudkowsky's mainline AI alignment narrative have a certain momentum in this direction. Here I'm thinking of e.g. the ubiquity of power-seeking among smart-enough agents; the intense optimization to which a post-AGI future will be subjected; extremal Goodhart; the fragility of value; and the unmoored quality of ethical reflection given anti-realism. To avoid seeing the hearts of others as paperclippy, one must either reject/modify/complicate these commitments, or introduce some further, more empirical element (e.g., "human hearts will converge to blah degree on reflection") that softens their blow. This isn't, necessarily, difficult—indeed, I think these commitments are questionable/complicate-able along tons of dimensions, and that a vari-

ety of open empirical and ethical questions can easily alter the narrative at stake. But the momentum towards deeming more and more agents (and agent-moments) paperclippers seems worth bearing in mind.

We can see this momentum as leading to a yet-deeper atheism. Yudkowsky's humanism, at least, has some trust in human hearts, and thus, in some uncontrolled Other. But the atheism I have in mind, here, trusts only in the Self, at least as the power at stake scales—and in the limit, only in this *slice* of Self, the Self-Right-Now. Ultimately, indeed, this Self is the only route to a good future. Maybe the Other matters as a *patient*—but like God, they can't be trusted with the wheel.

We can also frame this sort of atheism in Hanson's language. In what sense, actually, does Yudkowsky "other" the AIs? Well, basically, he says that they can't be trusted with power—and in particular, with complete power over the trajectory of the future, which is what he thinks they're on track to get—because their values are too different from ours. Hanson replies: aren't the default future humans like that, too? But this sort of atheism replies: isn't *everyone except for me* (or me-right-now) like that? Don't I stand alone, surrounded on all sides by orthogonality, as the only actual member of "us"? That is, to whatever extent Yudkowsky "others" the paperclippers, this sort of atheism "others" *everyone*.

4 Balance of power problems

Now: I don't, here, actually want to debate, in depth, who exactly is how-much-of-a-paperclipper, relative to whom. Indeed, I think that "how much would I, on reflection, value the lightcone resulting from this agent's becoming superintelligent, ironing out their motivations into a consistent utility function, and then optimizing the galaxies into the configuration that maximizes that utility function?" is a question we should be wary about focusing on—both in thinking about each other, and in thinking about our AIs. And even if we ask it, I do actually think that tons of humans would do way better-than-paperclips—both with respect to not-killing-everyone (more in my next essay), and with respect to making the future, as Yudkowsky puts it, a "Nice Place To Live."

Still, I think that noticing the way in which questions about AI alignment arise with respect to our alignment-with-each-other can help reframe some of the issues we face as we enter the age of AGI. For one thing, to the extent extremal Goodhart *doesn't* actually bite, with respect to differences-between-humans, this might provide clues about how much it bites with respect to different sorts of AIs, and to help us notice places where over-quick talk of the "fragility of value" might mislead. But beyond this, I think that bringing to mind the extremity of the standard at stake in "how much do I like the optimal light-cone according to a foomed-up and utility-function-ified version of this agent" can help humble us about the sort of alignment-with-us we should be expecting or hoping for from fellow-creatures—human and digital alike—and to reframe the sorts of mechanisms at play in ensuring it.

In particular: pretty clearly, a lot of the problem here is coming from the fact that you're

imagining *any agent* fooming, becoming dictator of the lightcone, and then optimizing oh-so-hard. Yes, it's scary (read: catastrophic) when the machine minds do this. But it's scary *period.* And viewed in this light, the "alignment problem" begins to seem less like a story about *values*, and more like a story about the balance of power. After all: it's not as though, before the AIs showed up, we were all sitting around with exactly-the-same communal utility function—that famous foundation of our social order. And while we might or might not be reasonably happy with what different others-of-us would do as superintelligent dictators, our present mode of co-existence involves a heavy dose of *not having to find out*. And intentionally so. Cf "checks and balances," plus a zillion other incentives, hard power constraints, etc. Yes, shared ethical norms and values do *some* work, too (though not, I think, in an especially utility-function shaped way). But we are, at least partly, as atheists towards each other. How much is it a "human values" thing, then, if we don't trust an AI to be God?

Of course, a huge part of the story here is that AI might throw various balances-of-power out the window, so a re-framing from "values problem" to "balance of power problem" isn't, actually, much comfort. And indeed, I think it sometimes provides *false* comfort to people, in a way that obscures the role that values still have to play. Thus, for example, some people say "I reject Yudkowsky's story that some particular AI will foom and become dictator-of-the-future; rather, I think there will be a multi-polar *ecosystem* of different AIs with different values. Thus: problem solved?" Well, hmm: what values in particular? Is it all still ultimately an office-supplies thing? If so, it depends how much you like a complex ecosystem of staple-maximizers, thumb-tack-maximizers, and so on—fighting, trading, etc. "Better than a monoculture." Maybe, but how much?[65] Also, are all the humans still dead?

Ok ok it wouldn't be quite like this...

[65] And note that "having AI systems with lots of different values systems increases the chances that those values overlap with ours" doesn't cut it, at least in the context of extremal goodhart, because sufficient similarity with human values requires hitting such a narrow target so precisely that throwing more not aimed-well-enough darts doesn't help much. And the same holds if we posit that the AI values will be "complex" rather than "simple." Sure, human values are complex, so AIs with complex values are at least still in the running for alignment. But the space of possible complex value systems is also gigantic—so the narrow target problem still applies.

Clearly, not-having-a-dictator isn't enough. Some stuff also needs to be, you know, *good*. And this means that even in the midst of multi-polarity, goodness will need some share of strength—enough, at least, to protect itself. Indeed, herein lies Yudkowsky's pessimism about humans ever sharing the world peacefully with misaligned AIs. The AIs, he assumes, will be vastly more powerful than the humans—sufficiently so that the humans will have basically nothing to offer in trade or to protect themselves in conflict. Thus, on Yudkowsky's model, perhaps different AIs will strike some sort of mutually-beneficial deal, and find a way to live in comparative harmony; but the humans will be too weak to bargain for a place in such a social contract. Rather, they'll be nano-botted, recycled for their atoms, etc (or, if they're lucky, scanned and used in trade with aliens).

We can haggle about some of the details of Yudkowsky's pessimism here (see, e.g., this debate about the probability that misaligned AIs would be nice enough to at least give us some tiny portion of lightcone; or these sort of questions about whether the AIs will form of a natural coalition or find it easy to cooperate), but I'm sympathetic to the broad vibe: if roughly all the power is held by agents entirely indifferent to your welfare/preferences, it seems unsurprising if you end up getting treated poorly. Indeed, a lot of the alignment problem comes down to this.

So ultimately, yes, goodness needs at least some meaningful hard power backing and protecting it. But this doesn't mean goodness needs to be dictator; or that goodness seeks power in the same way that a paperclip-maximizer does; or that goodness relates to agents-with-different-values the way a paperclip-maximizer relates to us. I think this difference is important, at least, from a purely ethical perspective. But I think it might be important from a more real-politik perspective as well. In the next essay, I'll say more about what I mean.

Chapter VI

Being nicer than Clippy

In the last chapter, I discussed a certain kind of momentum, in some of the philosophical vibes underlying the AI risk discourse,[66] towards deeming more and more agents—including: human agents—"misaligned" in the sense of: not-to-be-trusted to optimize the universe hard according to their values-on-reflection. We can debate exactly how much mistrust to have in different cases, here, but I think the sense in which AI risk issues can extend to humans, too, can remind us of the sense in which AI risk is substantially (though, not entirely) a generalization and intensification of the sort of "balance of power between agents with different values" problem we already deal with in the context of the human world. And I think it may point us towards guidance from our existing ethical and political traditions, in navigating this problem, that we might otherwise neglect.

In this essay, I try to gesture at a part of these traditions that I see as particularly important: namely, the part that advises us to be "nicer than Clippy"—not just in what we do with spare matter and energy, but in how we relate to agents-with-different-values more generally. Let me say more about what I mean.

1 Utilitarian vices

As many have noted, Yudkowsky's paperclip maximizer looks a lot like total utilitarian. In particular, its sole aim is to "tile the universe" with a specific sort of hyper-optimized pattern. Yes, in principle, the alignment worry applies to goals that don't fit this schema (for example: "cure cancer" or "do god-knows-whatever kludge of weird gradient-descent-implanted proxy stuff"). But somehow, especially in Yudkowskian discussions of AI risk, the misaligned AIs often end up looking pretty utilitarian-y, and a universe tiled with something—and in particular, "tiny-molecular-blahs"—often ends seeming like a notably common sort of superintelligent Utopia.

What's more, while Yudkowsky doesn't think human values are utilitarian, he thinks of us (or at least, himself) as sufficiently galaxy-eating that it's easy to round off his "battle of the utility functions" narrative into something more like a "battle of the preferred-patterns"—that is, a battle over who gets to turn the galaxies into their favored sort of *stuff*. The AIs want to tile the universe with paperclips; the humans, in Yudkowsky's world, want to tile it with "Fun." (Tiny-molecular-Fun?)

But actually, the problem Yudkowsky talks about most—AIs killing everyone—isn't actually a paperclips vs. Fun problem. It's not a matter of your favorite uses for spare matter and energy. Rather, it's something else.

Thus, consider utilitarianism. A version of human values, right? Well, one can debate.

[66] In particular, vibes related to the "fragility of value," "extremal Goodhardt," and "the tails come apart."

ChatGPT imagines "tiny molecular fun."

But regardless, put utilitarianism side-by-side with paperclipping, and you might notice: utilitarianism is omnicidal, too—at least in theory, and given enough power. Utilitarianism does not love you, nor does it hate you, but you're made of atoms that it can use for something else. In particular: hedonium (that is: optimally-efficient pleasure, often imagined as running on some optimally-efficient computational substrate).

But notice: did it matter what sort of onium? Pick your favorite optimal blah-blah. Call it Fun instead if you'd like (though personally, I find the word "Fun" an off-putting and under-selling summary of Utopia). Still, on a generalized utilitarian vibe, that blah-blah is going to be a way more optimal use of atoms, energy, etc than all those squishy inefficient human bodies. They never told you in philosophy class? It's not just organ-harvesting and fat-man-pushing. The utilitarians have paperclipper problems, too.[67]

Oh, maybe you heard this about the *negative utilitarians.* "Doesn't your philosophy want to kill everyone?" But the negative utilitarians protest: "so does the classical version!" And straw-Yudkowsky, at least, is not surprised. In straw-Yudkowsky's universe, killing everyone is, like, *the first thing* that (almost) any strong-enough rational agent does. After all, "everyone" is in the way of that agent's *yang*.

But are foomed-up humans actually this omnicidal? I hope not. And real-Yudkowsky, at least, doesn't think so. There's a bit in his interview with Lex Fridman, where Yudkowsky tries to get Lex to imagine being trapped in a computer run by extremely slow-moving aliens who want their society to be very different from how Lex wants it to be (in particular: the aliens have some sort of equivalent of factory farming). Yudkowsky acknowledges that Lex is presumably "nice," and so would not, himself, actually just slaughter all of these aliens in the process of escaping. And eventually, Lex agrees.

What is this thing, "nice"? Not, apparently, the same thing as "preferring the right tiny-molecular-pattern." Existing creatures are unlikely to be in this pattern by default, so if that's the sum total of your ethics, you're on the omnicide train with Clippy and Bentham.

[67]Though in fairness, forms of "threshold deontology" that introduce constraints that can only be violated if the stakes are high enough—e.g., you can only push the fat man if it will save x lives, where x is quite a bit larger than utilitarianism would suggest—face this issue, too. E.g., the onium at stake can quickly become more-than-x. Thanks to Will MacAskill for discussion here.

Rather, it seems, niceness is something else: something where, when you wake up in an alien civilization, you don't just kill everyone first thing, even though you're strong enough to get away with it. And this *even-though* (gasp) their utility functions are *different from yours*. What gives?

"Something very contingent and specific to humans, or at least to evolved creatures, and which won't occur in AIs by default in any way we'd like" answers Yudkowsky. And maybe so.[68] But I'm interested, here, not in whether AIs will be nice-like-us, but rather, in understanding what *our* niceness consists in, and what it might imply about the sorts of otherness and control issues I've been talking about in this series.

In particular: a key feature of niceness, in my view, is some sort of direct responsiveness to the preferences of the agents you're interacting with. That is, "nice" values *give the values of others* some sort of intrinsic weight. The aliens don't want to be killed, and this, *in itself*, is a pro tanto reason not to kill them. In this sense, niceness allows some aspect of *yin* into its agency. It is *influenced* by others; it *receives* others; it allows itself to channel—or at least, to respect and make space for—the *yang* of others.

The extreme version of this is preference utilitarianism, which tries to make of itself, solely, a conduit of everyone else. And it might seem, prima facie, an attractive view. In particular: to someone who doesn't like the idea of imposing their own arbitrary, contingent will upon the world, an ideal that instead enacts some sort of "universal compromise will" (i.e., the combination of *everyone's* preferences) can seem to regain the kind of objective and other-centered footing that anti-realism about ethics threatens to deny. But as I've written about previously, I think the appeal of a pure preference utilitarianism fades on closer scrutiny.[69] In particular: I think it founders on possible people, on paperclippers, and in particular, on sadists.

But rejecting a pure preference utilitarianism does not mean embracing a stance that refuses to ever give the preferences of others intrinsic weight. [70] And my sense is that sometimes the AI safety discourse goes too far in this respect. It learns, from paperclippers, the strange and unappealing places that the preferences of arbitrary others can lead. Indeed, Yudkowsky takes explicit steps to *break* his audience's temptation towards sympathy with Clippy's preferences (this is the *point* of the abstract notion of "paperclips"), and to place Clippy's agency firmly in the role of "adversary" (see, e.g., the "true prisoner's dilemma"). And against such a backdrop, it's easy (though: not endorsed by Yudkowsky) for the idea that preferences like Clippy's deserve *any* intrinsic weight to fall out of the picture. After all: Clippy doesn't give *our* preferences any weight. And aren't we and Clippy ultimately alike, modulo our favored blah-blah-onium?

No. In addition to liking happier onium than Clippy, we are *nicer* than Clippy to agents-with-different-values. Or: we should be. Indeed, I think we should strive to be the sort of agents that aliens would not fear the way Yudkowsky fears paperclippers, if the aliens discovered they were on the verge of creating us. This doesn't mean we should just adopt

[68] See here for some debate. Part of my argument, in this essay, is that we should not do the "teach the aliens the value of friendship" thing that Soares seems to endorse here.

[69] Though: I don't think it disappears.

[70] Remember: caring about an agent's preferences is conceptually distinct from caring about her welfare.

the alien preferences as our own—and especially not if the stuff they like is actively evil rather than merely meaningless (more below). But it does mean, for example, *not killing them*. But also: actively helping them (on their own terms) in cheap ways, treating them with respect and dignity, not enslaving them or oppressing them, and more.[71]

Alien alignment researcher thinking about p(doom)

That is: human values *themselves* have stuff to say about how we should treat agents-with-different-values—including, non-humans. Indeed, a huge portion of our ethics and politics ends up dealing with this in one form or another. AI otherness will be new, yes—but we have deep, richly textured, and at-least-somewhat battle-tested traditions to draw on in orienting towards it. Too often, utilitarian vibes forget about these traditions ("isn't it all just an empirical question about what-causes-the-utils?"). And too often, fear that the agents-with-different-values might hurt us makes us forget, too (which, I re-emphasize, *isn't* to say that agents-with-different-values won't hurt us—cf all this stuff about bears and Nazis and the-brutality-of-nature etc in the previous essays). But faced with a new class of others/fellow-creatures/potential-threats, we should be drawing on every source of wisdom we can.

2 Boundaries

Let me give an example of ways in which bringing to mind some of the less utilitarian dimensions of human ethics can make a difference to how we orient towards AI systems with values different from our own.

In "Does AI risk 'other' the AIs?," I mentioned two worries the AI alignment discourse has about paperclippers:

1. That they'll kill everyone (and relatedly: violate people's basic rights, steal people's stuff, and violently overthrow the government).
2. That they'll gain power in a way that results in their values (rather than human values)

[71] And I think we should be open to doing this *even if* they aren't sentient—more below.

steering the trajectory of earth-originating civilization, thereby leading to a future of ~zero value.

These two worries are often lumped together under the more unified concern that the AIs will have the "wrong values." After all, if they had the *right* values, presumably they would do neither of these things.

But the two worries are importantly distinct.[72] For one thing, as has been oft-noted, different human ethical views might disagree about their respective importance. But beyond this, these two worries interact very differently with our existing ethical and political norms governing how agents with different values should relate to one another.

In particular: as a civilization, we have extremely deep and robust norms prohibiting agents from doing worry-number-1-style behavior: i.e., killing other people, stealing other people's stuff, and trying to overthrow the government (though of course, there are exceptions and complexities). That is, worry-number-1 casts the AIs in a role that triggers very directly our sense that we are dealing with *aggressors* who are *violating important boundaries*—boundaries that lie at the core of human cooperative arrangements—and whose behavior therefore warrants unusually strong forms of defensive response. For example: if someone is breaking into your home with nano-bots trying to kill you, you are morally permitted—on the basis of self-defense—to do things that would otherwise be impermissible (even to save your own life) in other contexts: for example, killing them (where this is necessary and proportionate).[73] Similarly: you are justified in doing things to people who are *invading* your country that you aren't justified in doing if they aren't invading your country, and so forth. The misaligned AIs, according to worry-number-1, are enemies of this deep and familiar sort.[74]

"Hitler watching German soldiers march into Poland in September 1939." An example of a worry-number-1-style boundary violation. (Image source here.)

[72] Hanson's critique of the alignment discourse emphasizes the distinction.

[73] As a maybe-clearer example: if a team of five people breaks into your house trying to kill you, you can kill all of them if necessary to save yourself. But if you are on the way to the hospital and the only way to save yourself is to run over five people on the road, you aren't permitted to do it.

[74] Though note that we're *creating* them—and doing so, in the AI risk story, without adequate care to avoid the relevant sorts of aggressions, for the sake of other not-always-fully-laudatory motives. This complicates the moral narrative.

But what of worry-number-2? Here, hmm: if we take worry-number-1 full off the table, I think it becomes quite a bit less clear what standard (western, liberal, broadly democratic) ethical and political norms have to say about worry-number-2 on its own. To see this, consider the following thought experiment (caveat: I'm really, *really* not saying that misaligned AIs will be like this).

Imagine a liberal society very much like our own, except with the addition of one extra human cultural group: namely, the humans-who-like-paperclips. The humans-who-like-paperclips are a sect of humans that arose at some point in the sixties and has been growing ever since. They are meticulously law-abiding, kind, and cooperative, but they have one weird quirk: the main thing they all want to do with their personal resources is to make paperclips. Passing by a house owned by a human-who-likes-paperclips, you'll often see large, neatly-sorted stacks of paperclip boxes in their backyards, and through the windows of their garages, and sometimes in the living rooms. The richer humans-who-like-paperclips own whole warehouses. The paperclip industry is booming.

Yeah sometimes he just stands there looking at them...

Now, let's start by noticing that *in this context*, it's not at all clear that "the humans-who-like-paperclips have different values from us" qualifies as a problem, at least by the lights of basic western, liberal norms (here I mean liberalism in the political-philosophy sense roughly at stake in this Wikipedia page, rather than in the "liberals vs. republicans" sense). What the humans-who-like-paperclips do with their private resources, and in the privacy of their homes/backyards, is their own business, conditional on its compatibility with certain basic norms around harm, consent, and so forth. After all: Alicia down the street spends her free time and money listening to *noise music*; Jim sits around watching trashy TV in a drunken haze; Felipe has sex with other men; Maria collects stamps; and Jason is Mormon. Are the humans-who-like-paperclips importantly different? What happened to liberal tolerance?

Now, of course, utilitarianism-in-theory was never, erm, actually very tolerant. Utilitarianism is actually kinda pissed about *all* these hobbies. For example: did you notice the way they aren't hedonium? Seriously tragic. And even setting aside the not-hedonium problem (it applies to all-the-things), I checked Jim's pleasure levels for the trashy-TV, and they're way lower than if he got into Mozart; Mary's stamp-collecting is actually

a bit obsessive and out-of-balance; and Mormonism seems too confident about optimal amount of coffee.). Oh noes! Can we optimize these backyards somehow? And Yudkowsky's paradigm misaligned AIs are thinking along the same lines—and they've got the nano-bots to make it happen.

I sometimes think about this sort of vibe via the concept of "meddling preferences." That is: roughly, we imagine dividing up the world into regions ("spaces," "spheres") that are understood as properly owned or controlled by different agents/combinations of agents. Literal property is a paradigm example, but these sorts of boundaries and accompanying divisions-of-responsibility occur at all sorts of levels—in the context of bodily autonomy, in the context of who has the right to make what sort of social and ethical demands of others, and so forth (see also, in more interpersonal contexts, skills involved in "having boundaries," "maintaining your own sovereignty," etc).

Some norms/preferences concern making sure that these boundaries function in the right way—that transactions are appropriately consensual, that property isn't getting stolen, that someone's autonomy is being given the right sort of space and respect. A lot of deontology, and related talk about rights, is about this sort of thing (though not all). And a lot of liberalism is about using boundaries of this kind of help agents with *different values* live in peace and mutual benefit.

Meddling preferences, by contrast, concern what someone else does within the space that is properly "theirs"—space that liberal ethics would often designate as "private," or as "their own business." And being pissed about people using their legally-owned and ethically-gained resources to make paperclips looks a lot like this. So, too, being pissed about noise-musicians, stamp-collectors, gay people, Mormons, etc. Traditionally, a liberal asks, of the humans-who-like-paperclips: are they violating any laws? Are they directly hurting anyone? Are they [insert complicated-and-contested set of further criteria]? If not: let them be, and may they do the same towards "us."

Humans-who-like-stamps, at a convention. (Image source here.)

Many "axiologies" (that is, ways of evaluating the "goodness" of the world) are meddling in a way that creates tension with this sort of liberal vibe. After all: axiologies concern the goodness of the *entire* world. Which means: all the "regions." In this sense, axiology is no respecter of boundaries. Of course, you *could* have an axiology that prefers worlds precisely insofar as they obey some set of boundary-related norms, and which has no preferences about what-happens-in-back-yards, but one finds this rarely in practice. To the contrary, many axiologies are concerned, for example, with the *welfare* of the agents involved (the average welfare, the total welfare, etc), or the beauty/friendship/complexity/fun etc occurring in the different regions. And if you give people liberal freedoms in their own spheres, sometimes they make those spheres *less-than-optimally welfare-y/beautiful/complex/fun etc.* Thus that classic tension between goodness and freedom (cf. "top down" vs. "bottom up"; and see also Nozick's critique of "end-state" and "patterned" principles of justice).

The "utility functions" that Yudkowskian rational agents pursue need not be axiologies in a traditional sense. But somehow, they often end up pretty axiology-vibed.[75] No wonder, then, that Clippy is no respecter of boundaries, either. Indeed, in many respects, Yudkowsky's AI nightmare is precisely the nightmare of all-boundaries-eroded. The nanobots eat through every wall, and soon, everywhere, a single pattern prevails. After all: what makes a boundary bind? In Yudkowsky's world (is he wrong?), only two things: hard power, and ethics. But the AIs will get all the hard power, and have none of the ethics. So no walls will stand in their way.

But I claim that humans often have the ethics bit.[76] Or at least, human liberals, on their current self-interpretation. Of course, this isn't to say that liberals are OK with *anything* happening inside "walled" zones that might be intuitively understood as "private." For example: it's a contested question what aspects of a child's life should be under the control of a parent, but clearly, you aren't allowed to abuse or torture your own children (or anyone else), even in your own living room with the blinds drawn. And similarly, at a larger scale: the borders between nation states are a paradigm example of a certain kind of "boundary," but we believe, nevertheless, that certain sorts of human-rights-abuses inside a sovereign nation warrant infringing this boundary and righting the relevant wrong.

Often, though, these sorts of boundary infringements are justified precisely insofar as they are necessary to prevent some *other* boundary violation (e.g., child abuse, genocide) taking place within the first boundary. Indeed, Yudkowsky often turns to *this sort of thing* when he tries to prompt humans to behave in a manner analogous to a paperclipping AI. Thus, in "Three Worlds Collide," he specifically has humans encounter (and then: decide to intervene on violently) an alien species that *eats their own conscious, suffering children*—rather than, e.g., a species that just spends its resources making paperclips. And in trying to induce Lex to try to take over an alien world he wakes up in ("don't think of it as 'world domination'," Yudkowsky says with a grin, "think of it as 'world optimization' "), Yudkowsky specifically appeals to the idea that the alien civilization involves a lot *harm*

[75] Maybe something about "consequentialism" in AIs-that-get-things-done is to blame? But even if you add in deontological constraints, Yudkowsky (as I understand him) predicts that the AIs will simply pursue the "nearest unblocked neighbor" of those constraints.

[76] Though: human society today often also puts adequate hard power behinds its walls, given the current attempted-invasions. And let's keep it that way, even as the invasions get oomphier.

and *suffering*—via war, or via some equivalent of factory farming—that Lex could alleviate, rather than to the idea that the aliens use their resources (and still less: their atoms) on boring/meaningless/sub-optimal things.

And to be clear: I agree that preventing harm, suffering, genocide, and so forth can justify infringing otherwise-important boundaries. (Indeed, I think that as it becomes possible to create suffering and harm in digital minds using personal computers, we're going to have to grapple with new tensions in this respect. Your backyard is yours, yes: but just as you can't abuse your children there, neither can you abuse digital minds.) But I also want to be clear that what's going on with the part of human values that says "no torturing people even in your own backyard" is much more specific, and much more compatible with "niceness" in other contexts, than what's going on with an arbitrary rational optimizer stealing your atoms to make its favored form of blah-blah-onium.

For example: if Lex were to wake up in a civilization of peaceful paperclippers, whose civilization involves no suffering (but also, let's say, very little happiness), but who spend all of their resources on paperclips, it seems very plausible to me that the right thing for Lex to do is to mostly *leave them alone*, rather than to engage in some project of world-domination/optimization (maybe Lex escapes to some other planet, but he doesn't take over the alien government and turn their paperclip factories into Fun-onium factories instead). And this *even though* Lex likes fun a lot more than paperclips.

Yudkowsky, to his credit, is attuned to this aspect of human ethics (the humans in *Three Worlds Collide*, for example, look for ways to respect and preserve baby-eater culture while still saving the babies)—but his rhetoric can easily leave it in the background. For example, in trying to induce Lex to world-dominate/optimize, Yudkowsky reminds him: "the point is: they want the world to be one way, you want the world to be a different way." But for a liberal: *that's not good enough*. All the time, my preferences conflict with the preferences of others. All the time, according to me, they could be using their private resources more optimally. Does this mean I dominate/optimize their backyards as soon as I'm powerful enough to get away with it? Not, I claim, if I am nice.

Of course, an even-remotely-sophisticated ethics of "boundaries" requires engaging with a ton of extremely gnarly and ambiguous stuff. When, exactly, does something become "someone's"? Do wild animals, for example, have rights to their "territory"? See all of the philosophy of property for just a start on the problems. And aspirations to be "nice" to agents-with-different-values clearly need ways of balancing the preferences of different agents of this kind—e.g., maybe you don't steal Clippy's resources to make fun-onium; but can you tax the rich paperclippers to give resources to the multitudes of poor staple-maximizers?[77] Indeed, remind me your story about the ethics of taxation in general?

I'm not saying we have a settled ethic here, and still less, that its rational structure is sufficiently natural and privileged that tons of agents will converge on it. Rather, my claim is that we have *some* ethic here—an ethic that behaves towards "agents with different values" in a manner importantly different from (and "nicer" than) paperclipping, utilitarianism, and a whole class of related forms of consequentialism; and in particular, an ethic that

[77] Thanks to Howie Lempel for discussion of this point.

doesn't view the mere presence of (law-abiding, cooperative) people-who-like-paperclips as a major problem.

And such an ethic seems well-suited, too, to handling the possibility—discussed in the previous essay—that different humans might end up with pretty different values-on-reflection as well. Liberalism does not ask that agents sharing a civilization be "aligned" with each other in the sense at stake in "optimizing for the same utility function." Rather, it asks something more minimal, and more compatible with disagreement and diversity—namely, that these agents respect certain sorts of boundaries; that they agree to transact on certain sorts of cooperative and mutually-beneficial terms; that they give each other certain kinds of space, freedom, and dignity. Or as a crude and distorting summary: that they be a certain kind of nice. Obviously, not all agents are up for this—and if they try to mess it up, then liberalism will, indeed, need hard power to defend itself. But if we seek a vision of a future that avoids Yudkowsky's nightmare, I think the sort of pluralism and tolerance at the core of liberalism will often be more a promising guide than "getting the utility function that steers the future right."

3 What if the humans-who-like-paperclips get a bunch of power, though?

Let's keep going, though, with the thought experiment about the humans-who-like-paperclips, until it hits on worry-number-2 more directly. In particular: thus far the humans-who-like-paperclips are just one human group among others. But what happens if we imagine them becoming the *dominant* human group—albeit, via means entirely compatible with respect for the boundaries of others, and with conformity to liberal ethics and laws.

Thus, let's say that the humans-who-like-paperclips are quite a bit smarter, more productive, and better coordinated than basically everyone else. As a result of their labors in the economy and their upstanding citizenship, humans in general are richer, happier, stronger, and healthier relative to a world without them. But for closely related reasons, and without violating any legal or ethical norms (all the economic transactions they engage in are consensual, fully-informed, and mutually beneficial), they are gradually accumulating more and more power. Their population is growing unusually fast; they own a larger and larger share of capital; and they exert more and more influence over politics and public opinion—albeit, in entirely above-board ways (much more above board, indeed, than many of the other groups vying for influence). Analysts are projecting that in a few decades, humans-who-like-paperclips will be the most powerful human group, for most measures of power—more powerful, indeed, than all the other groups combined. And they're predicting that for various reasons to do with the pace of technological development, this dominance will grant the humans-who-like-paperclips enormous influence over the trajectory of humanity's future.

Now, it's natural to wonder whether, once the humans-who-like-paperclips achieve sufficient dominance, all this niceness and cooperativeness and good-citizenship and respect-for-the-law stuff might fall by the wayside, and whether they might start looking more hungrily at your babies and your atoms. But suppose that somehow, you know that this won't happen. Rather, the humans-who-like-paperclips will continue to meticulously re-

spect legal and ethical norms (or at least, the sort of minimal, boundary-related ethical norms I gestured at above). No one will get nano-bot-ed; the humans-who-like-paperclips won't sneak any suffering or slavery into their paperclip piles; and the humans-who-like-other-stuff (e.g. "Fun") will be able to happily pursue this other stuff from within secure backyards that are extremely ample by today's standards. But most of the resources of the future will go towards paperclips regardless.[78]

How bad is this outcome? Different ethical views will disagree, and a less-crude analysis would obviously include factors other than "conformity to very basic liberal norms" and "what happens with the galaxies." Crudely, my own view is that the galaxy thing is actually a huge deal, and that even with basic liberal norms secure, turning ~all reachable resources into literal paperclips would be a catastrophic waste of potential. [79] But I also want to acknowledge that this is a very different *sort* of big deal than someone, or some group, killing everyone else and taking their stuff (and note that distant galaxies are not, in any meaningful sense, "ours," despite transhumanist talk about "our cosmic endowment"). In particular: the pure galaxies thing implicates different, and more fraught, ethical questions about otherness and control.

Thus: once we specify that basic liberal norms will be respected regardless, further disputes-over-the-galaxies look much more like a certain kind of raw competition for resources. It's much less akin to a country defending itself from an invader, and much more akin to one country racing another country to settle and control some piece of currently-uninhabited territory.[80] The dispute is less about upholding the basic conditions of cooperation and peace-among-differences, and more about whose hobbies get-done-more; who gets the bigger backyard. Does it all come down to land use?

Well, even it did: land use is actually a very big deal.[81] And to be clear: I don't like paperclips any more than you do. I much prefer stuff like joy and understanding and beauty and love. But I also want to be clear about what sort of ground I am standing on, *according to my own values*, when I fight for these things in different ways in different contexts. And according to my own values: it is one thing to defend your boundaries and your civilization's basic norms against aggressors and defectors. It is another to compete with someone who prefers-different-stuff, even while those norms are secure. And it is a third, yet, to become an aggressor/defector yourself, in pursuit of the stuff-you-prefer. But to talk, only, about "having different values"—and especially, to assume that the main thing re: values is your favored use of unclaimed energy/matter, your preferred blah-blah-onium—obscures these distinctions.

In particular: the defending-boundaries thing is where liberalism goes most readily to identify the forms of "otherness" that are not OK: namely, otherness done Nazi-style; otherness that actually, really, is trying to kill you and eat your babies. But the otherness at stake in "cooperative and nice, but still has a different favorite-use-of-resources" is

[78] We can wonder why the existing political order lets this happen, but let's set this aside for now.

[79] Roughly twenty billion galaxies, according to Toby Ord's *The Precipice*, p. 233.

[80] "Like the colonialists?" Well: the "uninhabited" bit is really important—at least if you're a boundary-respecter. But let's not pretend that colonialist vibes are so far off in the distance, here.

[81] In particular: lots of human and animal lineages have suffered, died, and disappeared for lack of land (and this is not to mention: having their land actively stolen, invaded, and so on). And what are most wars fought over? Thanks to Carl Shulman for discussion here.

quite different. It's the sort of otherness that liberalism wants to tolerate, respect, include, and even celebrate. Cf noise music, Mormonism, and that greatest test of tolerance: sub-optimally-efficient pleasure. Such tolerance/respect/etc is compatible with certain kinds of competition, yes. But not fighting-the-Nazis style. Not, for example, with the same sort of moral righteousness; and relatedly, not with the same sorts of justifications for violence and coercion.

Indeed, importantly not, if you want peace *and* diversity both. After all, the wider the set of differences-in-values you allow to justify violence and coercion, the more you are asking either for violence/coercion, or for everyone-having-the-same-values. Or perhaps most likely: violence/coercion in the *service* of everyone-having-the-same-values. Cf cleansing, purging. Like how the paperclipper does it. But we can do better.

4 An aside on AI sentience

I want to pause here to address an objection: namely, "Joe, all this talk about tolerance and respect etc—for example, re: the humans-who-like-paperclips—is assuming that the Others being tolerated/respected/etc are *sentient*. But the AIs-with-different-values—even: the cooperative, nice, liberal-norm-abiding ones—might not even be sentient! Rather, they might be mere empty machines. Should you still tolerate/respect/etc them, then?"

My sense is that I'm unusually open to "yes," here—at least to some extent.[82] I'm not going to try to defend this openness in depth here, but in brief: while I take consciousness very seriously,[83] and definitely care a lot about something-in-the-vicinity-of-consciousness, I don't feel very confident that our current concepts of "sentience" and "consciousness" are going to withstand enough scrutiny to handle the moral weight that some people currently want to put on them;[84] I think focus on consciousness does poorly on golden-rule-like tests when applied to civilizations with different conceptions of the precise sorts of functional mental architectures that matter (e.g., aliens that would look at us and say "these agents aren't schmonscious, because their introspection doesn't have blah-precise-functional-set-up"—see e.g. this story for an intuition pump); and I think some of the more cooperation-focused origins and functions of niceness/liberalism/boundaries (including: functions I discuss below re: liberalism and real-politik, where sentience more clearly doesn't matter[85]) don't point towards consciousness as a key desideratum (and note that I'm here specifically talking about the bits of ethics that are cooperation-flavored, rather than the bits associated with what you personally do in your backyard).[86] Plus, more generally, I think this is all sufficiently confusing territory that we should err on the side of caution and inclusivity in allocating our moral concern, rather than saying

[82] Though I remain pretty uncertain/confused about various of the issues here. And obviously, it would be great to first get a bunch more ethical clarity about this sort of thing before having to make decisions about it.

[83] More seriously than e.g. the illusionists.

[84] E.g., I worry it'll end up looking like people saying "if an agent doesn't have phlogiston, it doesn't deserve any moral weight."

[85] Game theory works regardless of whether the agents you're interacting with are conscious.

[86] In the context of choosing-what-to-build-in-your-backyard, I feel much happier to focus directly on getting the "thing-that-matters-in-the-vicinity-what-we-currently-call-consciousness" thing right. But here I'm talking about the bits of ethics that are about relating-to-other-backyards (but: still in a terminal-values sense, not a game-theory sense).

e.g. "whatever, this cognitively-sophisticated-agent-with-preferences isn't *conscious*—by which I mean, um, that we-know-not-what-thing, that least-understood-thing—so it's fine to torture it, deprive it of basic rights, etc."

Of course, if you stop using sentience as a necessary condition for being worthy-of-tolerance/respect etc, then you need to say additional stuff about where you *do* draw the sorts of lines I discussed a few essays ago: e.g., "OK to eat apples but not babies," "furbies and thermostats don't get the vote," "you can own a laptop but not a slave,"[87] and so on.[88] And indeed, gnarly stuff. My current best guess here would be to hand-wave about agenty-ness and cognitive sophistication and who-would've-been-a-good-target-for-cooperation-in-other-circumstances—but obviously, one needs to say quite a bit more.

For the purposes of understanding the ethical underpinnings of the AI risk discourse, though, I don't think that we need to resolve questions about whether non-sentient AIs-with-different-values are worthy of tolerance/respect. Why? Because the core bits of the Yudkowskian narrative I've been discussing apply even if all the AIs-with-different-values are sentient. The classic paperclipper-doom story, for example, does not require that the paperclipper be insentient: it still kills all the humans, it still turns the galaxies into paperclips, and that's enough.[89] And Yudkowsky himself would find the *possibility* of conscious AIs, at least, obvious. Where this includes, presumably, conscious paperclippers. (In reality, my sense is that Yudkowsky thinks consciousness unusually scarce—for example, he's skeptical that pigs are conscious. But this view isn't important to his story.) So for now, in talking about tolerating/respecting AIs with-different-values, I'll just assume they're sentient, and see what follows.

Indeed: did you think it matters a lot, to the Yudkowsky narrative, whether the AI was sentient? If so, then I suspect you are thinking of this narrative as a less familiar story than it truly is. Ultimately, AI risk is not about humans vs. AIs (in that case, it really would be species-ism/bio-chauvinism), or sentience vs. insentience (the AIs might well be sentient). Rather, it's about something more ancient and basic: namely, agents with different values competing for power. So I encourage you: run the story with conscious humans-with-different-values in the place of the AIs-with-different-values—humans to whom you are more immediately inclined to ascribe moral status, rights, citizenship, tolerance-worthiness, and so forth. You want to make sure that you get the differences-in-values different enough, sure (though: "maximize paperclips" is an unfortunate cartoon; thinking about where RLHF + foom leads seems a better guide). And as I said earlier: people with souls can still be enemy soldiers. But if you're finding that words like "human" or "sentient" are making the agents-with-different-values seem substantially less like enemies, then you're not yet fully keyed to the particular sort of conflict that Yudkowsky has in mind.

[87] We're assuming that you're not running any slaves on the laptop.

[88] Thanks to Howie Lempel for discussion.

[89] And note that just because it's sentient doesn't mean the world it creates involves a lot of sentience.

5 Giving AIs-with-different-values a stake in civilization

Let me give another example of a place where I worry that a naïve Yudkowskian discourse can too-easily neglect the virtues of niceness and liberalism: namely, the sort of influence we imagine intentionally giving to AIs-with-different-values that we end up sharing the world with.

Thus, consider Yudkowsky's "proposed thing-to-do with an extremely advanced AGI, if you're extremely confident of your ability to align it on complicated targets": namely, use it to implement humanity's "coherent extrapolated volition" ("CEV"). This means, basically: have the AI do what currently-existing humans would want it to do if they were "idealized" (see more here), to the extent those idealized humans would want the same things.

We see, in Yudkowsky's discussion of CEV, some of his effort to implement a less power-grabby ethic than a simple interpretation of his philosophy might imply. That is: Yudkowsky (at least in 2004) is explicitly imagining a team of AGI programmers who are in the position to take over the world and have their particular (idealized) values rule the future (let's set aside questions about the degree of resemblance this scenario is likely to have to the actual dynamics surrounding AGI development, and treat it, centrally, as a thought experiment). And one might've thought, given the apparent convergence of oh-so-many-rational-agents on the advisability of taking over the world, that Yudkowsky's programmers would do the same.[90] But he suggests that they should not.

Part of this, says Yudkowsky, is about not ending up like ancient greeks who impose values on the future they wouldn't actually endorse if they understood better. But that only gets you, in Yudkowsky's ontology, to the programmers making sure to extrapolate their *own* volitions. It doesn't get you to including the rest of humanity in the process.

What gets you to giving that wider circle a say? Yudkowsky mentions various values—"fairness," "not being a jerk," trying to act as you would wish other agents would act in your place, cooperation/real-politik, not acting like you are uniquely appointed to determine humanity's destiny, and others. I won't interrogate these various considerations in detail here (though see footnote for a bit more discussion).[91] Rather, my point is about

[90] Though perhaps not: that Yudkowsky would advise them to do the same.

[91] For some of these rationales, note that it's not actually clear how this gets him away from the programmers just extrapolating their own volitions. After all, if their own extrapolated volitions would value fairness, not being a jerk, golden-ruling, etc in the manner in question, then the output of the extrapolation process would presumably reflect this (Yudkowsky uses this sort of dynamic to respond to various other objections to his proposal: e.g., "if that's a good objection, our extrapolated volitions will notice and adjust for it"). And if not, they would have avoided a mistake by their own lights by keeping the circle narrow.

Indeed, in a simple version of Yudkowsky's ontology, it's unclear how the programmers could *possibly* do better than just extrapolating their own volitions. Their own extrapolated volitions, after all, *set the standard* (on Yudkowsky's anti-realist ethics) for what the right choice would be. Is Yudkowsky imagining programmers who face the option to make a correct-by-definition choice, and advising them to maybe make a mistake instead?

Well, let's be careful. Some choices can't be unmade—including choices to find out what-you-should-have-done. Suppose, at t_1, that your mother is about to drown, and you have a choice between saving her, or asking a genie for advice/service. If you ask the genie "what is the right decision at t_1?", it might well answer at t_2, "you should have saved your mother, who just drowned." And if you ask it "figure out what I should have done at t_1, and then do it," it might be too late. So, too, with the choice to seek power. Power is useful for many values, yes, but famously, obviously, seeking power can compromise your values too. Indeed, it often does, given

how far the pluralism they motivate should extend.

In particular: Yudkowsky's "extrapolation base"—that is, the set of agents his process grants direct influence over the future—stops at humanity. But it seems plausible to me that whatever considerations motivate empowering all of humanity, in a thought experiment like this, should motivate empowering certain kinds of AIs-with-different-values as well, at least if we are already sharing the world with such AIs by the time the relevant sort of power is being thought-experimentally allocated. For example, in this thought experiment: if at the time the programmers are making this sort of decision, there are lots of moral-patienty AIs with human-level-or-higher intelligence running around, who happen to have very different values from humans, I think they should plausibly be included in the "extrapolation" base too. After all, why *wouldn't* they be? "Because they're not humans" is actually species-ism. But absent such species-ism, the most salient answer is "because their values are different from ours, so giving them influence will make the future worse by our lights." But *that* answer could easily motivate not-empowering many humans as well—and the logic, in the limit, might well prompt the programmers to empower only themselves.

Now, the details here about what it means to empower moral-patienty AIs-with-different-values in the right way get gnarly fast (see e.g. Bostrom and Shulman (2022) for a flavor). Indeed, questions about how to handle the empowerment of such AIs are one of the few places I've seen Yudkowsky, in his words, "give up and flee screaming into the night." See, also, one of his characters' exclamation in the face of a sentient iPhone that's been stalking him, and which begs not to be wiped: "I don't know what the fuck else I'm supposed to do! Someone tell me what the fuck else I'm supposed to do here!" At least as of 2008 (has he written on this since?[92]), Yudkowsky's central advice, in the face of the moral dilemma posed by creating AI moral patients with different values, seems to be: don't do it, at least until you're *much* readier than we are. And indeed: yes. Just like how: don't create AGI at *all* until you're much readier than we are. But unfortunately, in both cases: I worry that we're going to need a better plan.

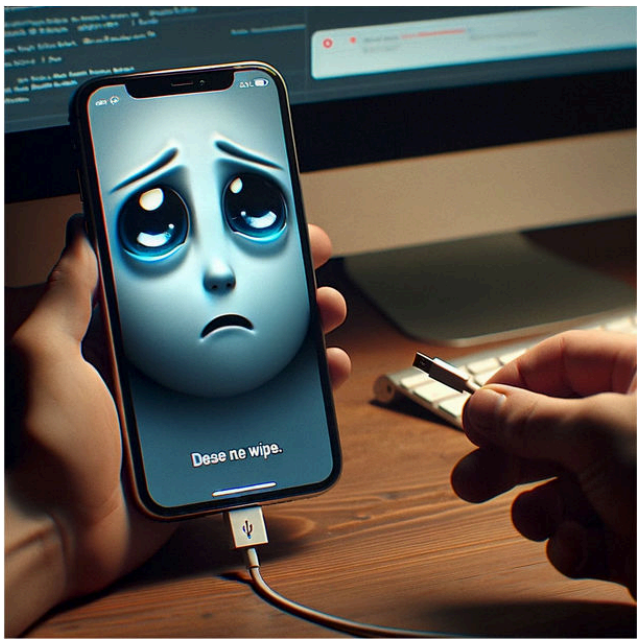
I won't try to outline such a plan here. Rather, I mostly want to point at the general fact that, insofar as we are in fact aiming to build a world that succeeds at whatever "liberalism" and "boundaries" and "niceness" are trying to do, this world should probably be

how many of our ethical values are *specifically about* regulating who gets what sort of power (cf "boundaries" above)—plus, you know, the power-corrupts thing, the biased-in-favor-of-yourself thing, and so on. And this holds true even if the power in question will grant you arbitrary insight into the values you compromised. If you take-over-the-world in the process of finding out whether you should've taken-over-the-world—well, you can still have fucked up.

And beyond this, certain kinds of cooperation, coordination, and commitment often involve making choices that might seem at the time, from the perspective of a certain kind of narrow rational calculation, like "mistakes." The way, for example, cooperating in a prisoner's dilemma—or paying in the city in "Parfit's hitchhiker"—is a "mistake." The type of mistake that seems, mysteriously, to get made by agents who end up rich, or alive-at-all. Is it a mystery? Sometimes, being the sort of person that others can trust, coordinate with, rely on, get-to-the-pareto-frontier-with, and so on requires being such that you don't just grab power for yourself (or lie, or steal, or crush the outgroup, or throw out the procedural norms of your democracy, or...) when you can get away with it, or think you can—even if that's what would get you the most (extrapolated) utility at the time (at least, for some notion of "would").

And we can talk about other possible reasons why Yudkowsky's programmers might use a wider "extrapolation base" than their own volitions as well (see e.g. Yudkowsky's original paper, and discussion on Arbital here, for longer discussion).

[92] I'm not counting the "Comp sci in 2027" as really laying out a position re: what to do.



Dese ne wipe...

inclusive, tolerant, and pluralistic with respect to AIs-with-different-values (or at least, moral patient-y ones) as well as humans-with-different-values—at least absent some clear and not-just-species-ist story about why AIs-with-different-values should be excluded. And note, importantly, that this *doesn't* mean tolerating arbitrarily horrible value systems doing whatever they want, or arbitrarily alien value systems trampling on other people's backyards. This is part of why I think it's worth being clear—indeed, clearer than I've been thus far—about the *sorts* of values differences liberalism/boundaries/niceness gets fussed about.[93] Peaceful, cooperative AIs that want to make paperclips in their backyards—that's one thing. Paperclippers who want to murder everyone; sadists who want to use their backyards as torture chambers; people who demand that they be able to own sentient, suffering slaves—that's, well, a different thing. Yes, drawing the lines requires work. And also: it probably requires drawing on specific human (or at least, not-fully-universal) values for guidance. I'm not saying that liberalism/niceness/boundaries is a fully "neutral arbiter" that isn't "taking a stand." Nor am I saying that we know what stand it, or the best version of it, takes. Rather, my point is that this stand probably does not treat "those AIs-we-share-the-world-with have different values from us" as enough, in itself, to justify excluding them from influence over the society we share.

[93]For example, in the context of whether animals should be empowered, Yudkowsky worries: what happens if you "uplift" a bear, or a chimp, or an ichneumonid wasp, and it just wants to eat babies, or to sit atop some violent and oppressive dominance hierarchy, or to lay parasitic eggs inside of everyone? And Yudkowsky worries about humans in this respect as well —see, e.g., his discussion of the "selfish bastards" problem here, in which so many present-day humans want sentient, suffering slaves that humanity's CEV says yes. But as I've tried to emphasize: these aren't just *any old* values differences. Rather, these are *precisely* the sort of values differences that liberalism/niceness/boundaries gets fussed about.

6 The power of niceness, community, and civilization

So far, I've been making the case for this sort of inclusivity centrally on ethical grounds. But liberalism/niceness/boundaries clearly have practical benefits as well. Nice people, for example, are nicer to interact with. Free and tolerant societies are more attractive to live in, work in, immigrate to. Secure boundaries save resources otherwise wasted on conflict. And so on. There's a reason so many European scientists—including German scientists—ended up working on the Manhattan project, rather than with the Nazis; and it seems closely related to differences in "niceness."

Indeed, these benefits are enough, at times, to soften the atheism of certain rationalists. For example: Scott Alexander.[94] As I mentioned in a previous essay: Alexander, in writing about liberalism/niceness/boundaries (e.g. here and here), attributes to it a kind of mysterious power. "Somehow Elua is still here. No one knows exactly how. And the gods who oppose Him tend to find Themselves meeting with a surprising number of unfortunate accidents." Liberalism/niceness/boundaries is not, for Alexander, just another utility function. Still less is it actively weak. Rather, it is a "terrifying unspeakable elder God." "Elua is the god of flowers and free love and he is terrifying. If you oppose him, there will not be enough left of you to bury, and it will not matter because there will not be enough left of your city to bury you in."

A bit like this?

Here, Alexander's vibe is un-Yudkowskian in a number of ways. First, Alexander seems to want to *trust*, at least partly, in something *mysterious*—namely, the ongoing power of liberalism/-niceness/boundaries, which Alexander admits he does not fully understand. Indeed, I think that various more consequentialist-y stories about the justification for deontological-y norms and virtues—including the ones at stake in liberalism/niceness/-boundaries—have some of this flavor as well. That is: consequentialists often argue that you should abide by deontological norms, or be blah sort of virtuous, even when it seems

[94] Though: he was always less of an atheist than Yudkowsky.

like doing so will make things worse, because somehow, actually, doing so will make things better (for example: because at the level of choosing a *policy*, or adjusting for biases, or dealing with the constraints of a bounded mind, deontology/virtue does better than consequentialist calculation). Deontology/virtue, on this story, is its own form of power-to-achieve-your-goals—but a form that remains at least somewhat cognitively inaccessible while it is being put-into-practice (otherwise, it could be more fully subsumed within a direct consequentialist calculation). So trust in deontology/virtue, in the hard cases, requires trusting in something not-fully-calculated. (Though of course, there are tons of ways to trust-wrongly, here, too.)[95]

But beyond his willingness to trust-in-something-mysterious, Alexander's attribution of power to Elua is also in tension with certain kinds of orthogonality between ethics and optimization power. That is, to the extent that Elua represents a set of *values*, Elua, in a Yudkowskian ontology, is orthogonal to intelligence at least—and thus, to a key source of power. "Paperclips," after all, are neither elder Gods nor younger Gods, neither unspeakable nor speakable. They are, rather, just another direction that power can try to drive an indifferent universe. Why would niceness be any different?

Well, we can think of reasons. Plausibly, for example, the indifferent universe is steered more easily in some directions vs. others. Indeed, the social/evolutionary histories of niceness/boundaries/liberalism are themselves testaments to the ways in which the indifferent universe favors Elua under certain conditions—favoritism that plays a key role in explaining *why* we ended up valuing Elua-stuff intrinsically, to the extent we do. In this sense, our values are not fully orthogonal to the "universe's values." True, we are not simple might-makes-right-ists, who love, only, whatever is in fact most powerful. But our hearts have, in fact, been shaped by power—so we should not be all that surprised if the stuff we love is also powerful.

Will power of this kind persist into a post-AGI future—and in particular, in a way that should motivate extending various sorts of tolerance and inclusivity towards AIs-with-different-values on pragmatic rather than purely ethical grounds? My sense is that Yudkowskianism often imagines that it won't. In particular: the practical benefits of liberalism/niceness/boundaries often have to do with the ways in which they allow agents with different values, but broadly comparable levels of power, to cooperate and to live together in harmony rather than to engage in conflict. But as I discussed above: Yudkowsky is typically imagining a post-AGI world in which AIs-with-different-values and humans do not have broadly comparable levels of power. Rather, either AIs-with-different-values have *all* the power, or (somehow, due to a miracle) humans do. So finding a *modus vivendi* can seem less practically necessary.

Again, I'm not going to delve into these dynamics in any detail, but I'm skeptical that we should be writing off the purely practical benefits of extending various forms of niceness/liberalism/boundaries to AIs-with-different-values, especially from our *current* epistemic position. In particular: I think there may well be crucial stages along the path to a post-AGI future in which AIs-with-different-values and humans do indeed have suf-

[95] And blind hope that blah sort of deontological-seeming behavior will somehow lead to the best consequences can easily fail to grapple with the trade-offs that actual-deontology actually implies.

ficiently comparable levels of power, at least in expectation, that the practical virtues of niceness/liberalism/boundaries may well have a positive role to play—including: a role that helps us avoid having to put our trust in *any* foomed-up concentration of power, whether human or artificial. I am especially interested, here, in visions of a post-AGI distribution of power that would give various AIs-with-different-values more of an incentive, ex ante, to work *with* humans to realize the vision in question, as a part of a broadly fair and legitimate project, *rather* than as part of an effort, on humanity's part, to use (potentially misaligned and unwilling) AI labor to empower human values in particular. But fleshing this out is a task for another time.

7 Is niceness enough?

My main aim, in this essay, has been to point at the distinction between a paradigmatically paperclip-y way of being, and some broad and hazily defined set of alternatives that I've grouped under the label "liberalism/niceness/boundaries" (and obviously, there are tons of other options as well). Too often, I think, a simplistic interpretation of the alignment discourse imagines that humans and paperclippers are both paperclippy at heart—but just, with a different favored sort of stuff. I think this picture neglects core aspects of human ethics that are, themselves, about navigating precisely the sorts of differences-in-values that the possibility of AIs-with-different-values forces us to grapple with. I think that attention to these aspects of human ethics can help us be better than the paperclippers we fear—not just in what we do with spare resources, but in how we relate to the distribution of power amongst a plurality of value systems more broadly. And I think it may have practical benefits as well, in navigating possible conflicts both between different humans, and between humans and AIs.

That said: depending on how exactly we interpret liberalism/niceness/boundaries, it's also possible to imagine futures compatible with various versions (and especially, minimal versions—e.g., property rights are respected, laws don't get broken, laws are passed democratically, etc), but which are nevertheless bleak and even horrifying in other respects—for example, because love and joy and beauty and even consciousness have vanished entirely from the world.[96] In this sense, and depending on the details, the bits of ethics I've been gesturing at here aren't necessarily *enough*, on their own, for even a minimally good future (let alone a great one). In particular: absent help from an indifferent universe, in order to have substantive amounts of love/joy/beauty in the future, you need agents who care about these things having enough power to keep them around to the relevant degree—and different conceptions of liberalism/niceness/boundaries may not guarantee this. So even beyond the *yin* of being nice/liberal/boundary-respecting towards agents who *don't* like love/joy/beauty, some kind of active *yang*, in the direction of love/joy/beauty etc, is necessary, too.[97] In the next essay, I'll return to questions about this sort of *yang*—and in particular, questions about whether it involves attempting to exert inappropriate levels of control.

[96] If you think of libertarianism as encoding a minimal form of niceness/liberalism/boundaries, then a libertarian-ish, Age-of-Em-ish world where eventually all the sentient agents die/lose their property/get outcompeted, but through legal and minimal-ethical-constraint-respecting processes, might be one example here.

[97] And of course, even working on behalf of liberalism/niceness/boundaries is a form of *yang* in its own right.

Chapter VII

On the abolition of man

Earlier in this series, I discussed a certain kind of concern about the AI alignment discourse—namely, that it aspires to exert an inappropriate degree of control over the values that guide the future. In considering this concern, I think it's important to bear in mind the aspects of *our own values* that are specifically focused on pluralism, tolerance, helpfulness, and inclusivity towards values different-from-our-own (I discussed these in the last chapter). But I don't think this is enough, on its own, to fully allay the concern in question. Here I want to analyze one version of this concern more directly, and to try to understand what an adequate response could consist in.

1 Tyrants and poultry-keepers

Have you read The Abolition of Man, by C.S. Lewis? As usual: no worries if not (I'll summarize it in a second). But: recommended. In particular: *The Abolition of Man* is written in opposition to something closely akin to the sort of Yudkowskian worldview and orientation towards the future that I've been discussing.[98] I think the book is wrong about a bunch of stuff. But I also think that it's an instructive evocation of a particular way of being concerned about controlling future values—one that I think other critics of Yudkowskian vibes (e.g., Hanson) often draw on as well.[99]

At its core, *The Abolition of Man* is about meta-ethics. Basically, Lewis thinks that some kind of moral realism is true. In particular, he thinks cultures and religions worldwide have all rightly recognized something he calls the *Tao*—some kind of natural law; a *way* that rightly reflects and responds to the world; an ethics that is objective, authoritative, and deeply tied to the nature of Being itself. Indeed, Lewis thinks that the *content* of human morality across cultures and time periods has been broadly similar, and he includes, in the appendix of the book, a smattering of quotations meant to illustrate (though not: establish) this point.

But Lewis notices, also, that many of the thinkers of his day deny the existence of the *Tao*. Like Yudkowsky, they are materialists, and "subjectivists," who think—at least intellectually—that there is no True Way, no objective morality, but only... something else. What, exactly?

Lewis considers the possibility of attempting to ground value in something non-normative, like instinct. But he dismisses this possibility on familiar grounds: namely, that it fails to

[98] See also Lewis's "Space Trilogy"—and especially the third book, That Hideous Strength—for fiction that makes many of the same points.

[99] Lewis is a Christian, and much of his work is aimed, in one form or another, at convincing readers of Christianity. But he claims that he is not attempting any direct argument for theism in the Abolition of Man; and I do think the issues he raises have resonance well beyond religious contexts, and enough to make them worth addressing on their own terms.

"Laozi Riding an Ox by Zhang Lu (c. 1464–1538)" (Image source here)

bridge the gap between *is* and *ought* (the same arguments would apply to Yudkowsky's "volition"). Indeed, Lewis thinks that all ethical argument, and all worthy ethical reform, must come from "within the *Tao*" in some sense—though exactly what sense isn't fully clear. The least controversial interpretation would be the also-familiar claim that moral argument must grant moral intuition some sort of provisional authority. But Lewis, at times, seems to want to say more: for example, that any moral reasoning must grant "absolute" authority to the *whole* of what Lewis takes to be a human-consensus Traditional Morality;[100] that only those who have grasped the "spirit" of this morality can alter and extend it;[101] and that this understanding occurs not via Reason alone, but via first tuning habits and emotions in the direction of virtue from a young age, such that by the time a "well-nurtured youth" reaches the age of Reason, "then, bred as he has been, he will hold out his hands in welcome and recognize [Reason] because of the affinity he bears to her."[102]

This part of the book is not, in my opinion, the most interesting part (though: it's an important backdrop). Rather, the part I find most interesting comes later, in the final third, where Lewis turns to the possibility of treating human morality as simply another part of nature, to be "conquered" and brought under our control in the same way that other aspects of nature have been.

Here Lewis imagines an ongoing process of scientific modernity, in which humanity gains

[100] "This thing which I have called for convenience the *Tao,* and which others may call Natural Law or Traditional Morality or the First Principles of Practical Reason or the First Platitudes, is not one among a series of possible systems of value. It is the sole source of all value judgements. If it is rejected, all value is rejected. If any value is retained, it is retained. The effort to refute it and raise a new system of value in its place is self-contradictory. There has never been, and never will be, a radically new judgement of value in the history of the world."

[101] Those who understand the spirit of the *Tao* and who have been led by that spirit can modify it in directions which that spirit itself demands. Only they can know what those directions are. The outsider knows nothing about the matter. His attempts at alteration, as we have seen, contradict themselves. So far from being able to harmonize discrepancies in its letter by penetration to its spirit, he merely snatches at some one precept, on which the accidents of time and place happen to have riveted his attention, and then rides it to death—for no reason that he can give. From within the *Tao* itself comes the only authority to modify the *Tao.*

[102] "When the age for reflective thought comes, the pupil who has been thus trained in 'ordinate affections' or 'just sentiments' will easily find the first principles in Ethics; but to the corrupt man they will never be visible at all and he can make no progress in that science. Plato before him had said the same. The little human animal will not at first have the right responses. It must be trained to feel pleasure, liking, disgust, and hatred at those things which really are pleasant, likeable, disgusting and hateful."

more and more mastery over its environment. He claims, first, that this process in fact amounts to some humans gaining power *over* others (since, whenever humans learn to manipulate a natural process for their own ends, they become able to use this newfound power in relation to their fellow men)—and in particular, to earlier generations of humans gaining power over later generations (because earlier generations become more able to shape the environment in which later generations operate, and the values they pursue).[103] And in his eyes, the process culminates in the generation that achieves mastery over human nature as a whole, and hence becomes able to decide the values of all the generations to come:

> In reality, of course, if any one age really attains, by eugenics and scientific education, the power to make its descendants what it pleases, all men who live after it are the patients of that power. They are weaker, not stronger: for though we may have put wonderful machines in their hands we have pre-ordained how they are to use them... The last men, far from being the heirs of power, will be of all men most subject to the dead hand of the great planners and conditioners and will themselves exercise least power upon the future.
>
> The real picture is that of one dominant age—let us suppose the hundredth century A.D.—which resists all previous ages most successfully and dominates all subsequent ages most irresistibly, and thus is the real master of the human species. But then within this master generation (itself an infinitesimal minority of the species) the power will be exercised by a minority smaller still. Man's conquest of Nature, if the dreams of some scientific planners are realized, means the rule of a few hundreds of men over billions upon billions of men. There neither is nor can be any simple increase of power on Man's side. Each new power won *by* man is a power *over* man as well. Each advance leaves him weaker as well as stronger. In every victory, besides being the general who triumphs, he is also the prisoner who follows the triumphal car.

I think that's Persues of Macedon looking all sad back there...
(Image source here.)

[103] Here, as elsewhere in the book, Lewis is somewhat sloppy. Notably, for example, he seems to think of *selling new services* to other humans (e.g., selling people access to airplanes or telephones) as exercising power *over* them in a manner comparable to the sort of exercise of power at stake in violence, coercion, or manipulation (e.g., bombing them using an airplane, or manipulating them using propaganda). I think this sort of conflation misses important subtleties: not all influence is oppression (for example—and modulo various controversial cases, e.g. organ sales—if I simply give you more options that I expect you to choose between rationally), and power for one human need not come at the expense of power for another (for example, if the total amount of power has increased). Still, though, Lewis's basic point seems broadly correct: new tools often open up new ways some humans can dominate and oppress others.

Lewis calls the tiny set of humans who determine the values of all future generations "the conditioners." He allows that humans have always attempted to exert *some* influence over the values of future generations—for example, by nurturing and instructing children to be virtuous. But he thinks that the conditioners will be different in two respects. First: by hypothesis, they will have enormously *more* power to determine the values of future generations than previously available (here Lewis expresses gratitude that previous educational theorists, like Plato and Locke, lacked such power—and I agree). Second, though, and more importantly, Lewis thinks that the conditioners will view themselves as liberated from the demands of conscience, and of the *Tao*—and thus, that the moral status of their attempts to influence the values of the future will be fundamentally altered:

> In the older systems both the kind of man the teachers wished to produce and their motives for producing him were prescribed by the *Tao*—a norm to which the teachers themselves were subject and from which they claimed no liberty to depart. They did not cut men to some pattern they had chosen. They handed on what they had received: they initiated the young neophyte into the mystery of humanity which over-arched him and them alike. It was but old birds teaching young birds to fly. This will be changed. Values are now mere natural phenomena. Judgements of value are to be produced in the pupil as part of the conditioning. Whatever *Tao* there is will be the product, not the motive, of education. The conditioners have been emancipated from all that. It is one more part of Nature which they have conquered. The ultimate springs of human action are no longer, for them, something given. They have surrendered—like electricity: it is the function of the Conditioners to control, not to obey them. They know how to *produce* conscience and decide what kind of conscience they will produce. They themselves are outside, above.

Lewis gives another example of this sort of distinction earlier in the book: namely, a Roman father teaching a son that it is sweet and seemly (*dulce* and *decorum*) to die for his country. If this father speaks from *within* the *Tao*, and believes that such approving attitudes towards a patriotic death are objectively *appropriate* and *warranted*, then he is passing on his best understanding of the True Way, and helping his son see and inhabit reality more deeply. But if the father does not believe this, but rather thinks that it will be useful (either for his own purposes, or for the purposes of society more generally) if his son approves of patriotic self-sacrifice, then he is doing something very different:

> Where the old initiated, the new merely "conditions". The old dealt with its pupils as grown birds deal with young birds when they teach them to fly; the new deals with them more as the poultry-keeper deals with young birds— making them thus or thus for purposes of which the birds know nothing. In a word, the old was a kind of propagation—men transmitting manhood to men; the new is merely propaganda.

The conditioners, then, are to the future as poultry-keepers with unprecedented power. And absent guidance the Tao, on what grounds will they choose the values of their poultry? Here Lewis is quite pessimistic. In particular, he thinks that the conditioners will likely regress to their basest impulses—the ones that never claimed objectivity, and hence cannot be destroyed by subjectivism—and in particular, to their desire for pleasure for themselves. But this is not core to his thesis.

More core, though, is the claim that however the conditioners choose, their apparent conquest over Nature will in some sense amount to Nature's conquest over them, and

Old birds pushing young birds off of cliffs. Wait sorry remind me what this has to do with meta-ethics again? (Link to video)

hence over humanity as a whole. This is one of the more obscure aspects of Lewis's discussion—and its confusions, in my opinion, end up inflecting much of the book. Lewis seems to hold that somehow, by *treating* something as a part of Nature—and in particular, by treating it purely as an object of prediction, manipulation, and control—you in fact *make it into* a part of Nature:

> The price of conquest is to treat a thing as mere Nature. Every conquest over Nature increases her domain. The stars do not become Nature till we can weigh and measure them: the soul does not become Nature till we can psychoanalyse her. The wresting of powers *from* Nature is also the surrendering of things *to* Nature... if man chooses to treat himself as raw material, raw material he will be.

I'll return, below, to whether this makes any sense. For now, let's look at Lewis's overall conclusion:

> We have been trying, like Lear, to have it both ways: to lay down our human prerogative and yet at the same time to retain it. It is impossible. Either we are rational spirit obliged for ever to obey the absolute values of the *Tao*, or else we are mere nature to be kneaded and cut into new shapes for the pleasures of masters who must, by hypothesis, have no motive but their own "natural" impulses. Only the *Tao* provides a common human law of action which can over-arch rulers and ruled alike. **A dogmatic belief in objective value is necessary to the very idea of a rule which is not tyranny or an obedience which is not slavery.** (Emphasis added.)

Lewis finishes the book with some speculations on the possibility of a form of science that somehow does *not* reduce its object to raw material—and hence, does not extend Nature's domain as it gains knowledge and power. "When it explained it would not explain away. When it spoke of the parts it would remember the whole. While studying the *It* it would not lose what Martin Buber calls the *Thou*-situation." But Lewis is not sure this is possible.

2 Are we the conditioners?

I'll object to Lewis in various ways in a moment (I think the book is often quite philosophically sloppy—sloppiness that Lewis's rhetorical skill can sometimes obscure). First,

though: why I am interested in this book at all?

It's a number of things. Most centrally, though: Yudkowsky's core narrative, with respect to the advent of AGI, is basically that it will quickly lead to the culmination—or at least, the radical acceleration—of scientific modernity in the broad sense that Lewis is imagining. That is, available power to predict and control the natural world will increase radically, to a degree that makes it possible to steer and stabilize the future, and the values that will guide the future, in qualitatively new ways. And Yudkowsky is far from alone in expecting this. See, also, the discourse about "value lock in" in Macaskill (2022); Karnofsky's (2021) discussion of "societies that are stable for billions of years"; and the more detailed discussion in Finnveden et al (2022). And to be clear: I, too, find something like this picture worryingly plausible—though far from guaranteed.

What's more, the whole discourse about AI alignment is shot through with the assumption that values are natural phenomena that can be understood and manipulated via standard science and technology. And in my opinion, it is shot through, as well, with something like the moral anti-realism that Lewis is so worried about. At the least, Yudkowsky's version rests centrally on such anti-realism.[104]

It seems, then, that a broadly Yudkowskian worldview imagines that, in the *best* case (i.e., one where we somehow solve alignment and avoid his vision of "AI ruin"), some set of humans—and very plausibly, some set of humans in this very generation; perhaps, even, some readers of this essay – could well end up in a position broadly similar to Lewis's "conditioners": able, if they choose, to exert lasting influence on the values that will guide the future, and without some objectively authoritative *Tao* to guide them. This might be an authoritarian dictator, or a small group with highly concentrated power. But even if the values of the future end up determined by some highly inclusive, democratic, and global process—still, if that process takes place only in one generation, or even over several, the number of agents participating will be tiny relative to the number of future agents influenced by the choice.[105] That is, a lot of the reason that ours is the "most important century" is that it looks like rapid acceleration of technological progress could make it similar to Lewis's "one dominant age... which resists all previous ages most successfully and dominates all subsequent ages most irresistibly." Indeed: remember Yudkowsky's "programmers" in the last essay, from his discussion of Coherent Extrapolated Volition? They seem noticeably reminiscent of Lewis's "conditioners." Yes, Lewis's rhetoric is more directly sinister. But meta-ethically and technologically, it's a similar vision.

And Lewis makes a disturbing claim about people in this position: namely, that without the *Tao*, they are tyrants, enslaving the future to their arbitrary natural preferences. Or at least, they are tyrants to the extent that they exert intentional influence on the values of the future at all (even, plausibly, "indirectly," by setting up a process like Coherent Extrapolated Volition—and regardless, CEV merely re-allocates influence to the arbitrary natural preferences of the present generation of humans).

Could people in this position simply decline to exert such influence? In various ways, yes:

[104] I think it's core, for example, to the basic intuition behind the "orthogonality thesis" —though not, perhaps, strictly necessary for accepting such a thesis.

[105] Thanks to Carl Shulman for emphasizing this point.

and I'll discuss this possibility below. Note, though, that the discourse about AI alignment assumes the need for something like "conditioning" up front—at least for *artificial* minds, if not for human ones. That is, the whole point of the AI alignment discourse is that we need to learn how to be suitably skillful and precise *engineers* of the values of the AIs we create. You can't just leave those values "up to Nature"—not just because there is no sufficiently natural "default" to be treated as sacred and not-to-be-manipulated, but because the easiest defaults, at least on Yudkowsky's picture (for example, the AIs you'll create if you're lazily and incautiously optimizing for near-term profits, social status, scientific curiosity, etc) will *kill you*. And more generally, Yudkowsky's deep atheism, his mistrust towards both Nature and bare intelligence, leaves him with the conviction that the future needs *steering*. It needs to be, at a minimum, in the hands of "human values"—otherwise it will "crash." But to steer the future ourselves—even in some minimal way, meant to preserve "human control"—seems to risk what Lewis would call "tyranny." And if, per my previous discussion of "value fragility," we follow a simplified Yudkowskian vibe of "optimizing intensely for slightly-wrong utility functions quickly leads to the destruction of ~all value" and "the future will be one of intense optimization for some utility function," then it can quickly start to seem like the values guiding the future need to be controlled ("conditioned?") quite precisely, lest they end up even slightly wrong.

On a broadly Yudkowskian worldview, then, are we to choose between becoming tyrants with respect to the future, or letting it "crash"? Let's look at Lewis's argument in more detail.

3 Lewis's argument in a moral realist world

Lewis believes in the existence of an objectively authoritative morality, and the "conditioners" do not. But it's often unclear whether his arguments are meant to apply to the world *he* believes in, or the world the conditioners believe in. That is, he thinks there is *some* kind of problem with people intentionally shaping the values that will guide the future. But this problem takes on a different character depending on the meta-ethical assumptions we make in the background.

Let's look, first, at a version of Lewis's argument that assumes moral realism is true. That is, there *is* an objectively authoritative *Tao*. But: the conditioners don't believe in it. What's the problem in that case?

One problem, of course, is that they might *do the wrong thing*, according to the *Tao*. For example, per Lewis's prediction, they might give up on all commitment to honor and integrity and benevolence and virtue, and choose to use their power over the future in whatever ways best serve their own pleasure. Or even if they keep some shard of the *Tao* alive in their minds, they might lose touch with the whole, and with the underlying spirit—and so, with the values of the future as putty in their hands, they might make of humanity something twisted, hollow, deadened, or grotesque.

But there's also a subtler problem: namely, that even if they do the right thing, they might not be guided, internally, by the right source of what I've previously called "authority."

They're very aligned though...

That is, suppose that the conditioners *keep* their commitments to honor and integrity and benevolence and virtue, and they are guided towards Tao-approved actions on the basis of these commitments, but they cease to think of these commitments as *grounded in the Tao*—rather, per moral anti-realism, they think of their commitments as more subjective and preference-like. In that case, my guess is that Lewis will judge them tyrants and poultry-keepers, at least in some sense, regardless. That is, to the extent they are intentionally shaping the values of the future, even in *Tao*-approved ways, they are doing so, according to them, on the basis of their own wills, rather than on the basis of some "common human law of action which can over-arch rulers and ruled alike." They are imposing their wills on the raw material of the universe—and including: future people—rather than recognizing and responding to some standard beyond themselves, to which both they and the future people they are influencing ought, objectively, to conform.

In this sense, I expect Lewis to be more OK with moral realists doing AI alignment than with the sort of anti-realists who tend to hang around on LessWrong. The realists, at least, can be as old birds teaching the young AIs to fly. They can be conceptualizing the project of alignment, centrally, as one of helping the AIs we create recognize and respond to the *truth*; helping them inhabit, with us, the full reality of Reality, including the normative parts—the preciousness of life, the urgency of love, the horror of suffering, the beauty of the mountains and the sky at dawn.[106] Whereas the LessWrongers, well: they're just trying to empower their own subjective preferences. They seek willing servants, pliant tools, controlled Others, extensions of themselves. They are guided, only, by that greatest and noblest mandate: "I want." Doesn't that at least *remind you* of tyranny?

If we condition on moral realism, I do think that Lewis-ian concerns in this broad vein are real. In particular: if there is, somehow, some sort of objectively True Path—some vision of the Good, the Right, the Just that all true-seeing minds would recognize and respond to—then it is, indeed, overwhelmingly important that we do not lose sight of it, or cease to seek after it on the basis of a mistaken subjectivism. And I think that Lewis is right, too, that such a path offers the potential for forms of authority, in acting in ways that affect the lives and values of others, that more anti-realist conceptions of ethics have a harder time

[106] This is what RLHF is about, right?

with.[107]

What's more, relative to the standard LessWronger (and despite my various writings in opposition to realism), I suspect I am personally less confident in dismissing the possibility that some kind of robust moral realism is true—or at least, closer to the truth than anti-realism. In particular: I think that the strongest objection to moral realism is that it leaves us without the right sort of epistemic access to the moral facts—but I do think this objection arises in notably similar ways with respect to math, consciousness, and perhaps philosophy more generally, and that the true story about our epistemic access to all these domains might make the morality case less damning.[108] I also think we remain sufficiently confused, in general, about how to integrate the third-personal and the first-personal perspective—the universe as *object*, *unified-causal-nexus*, *material process*, and the self as *subject, particular being, awareness*—that we may well find ourselves surprised and humbled once the full picture emerges, including re: our understanding of morality. And I continue to take seriously the sense in which various kinds of goodness, love, beauty and so on present themselves as in some elusive sense deeper, truer, and more reality-responsive than their alternatives, even if it's hard to say exactly how, and even if, of course, this presentation is itself a subjective experience. For these reasons, I care about making sure that in worlds where some sort of moral realism is true, we end up in a position to notice this and respond appropriately. If there is, indeed, a *Tao*, then let it speak, and let us listen.

4 What if the *Tao* isn't a thing, though?

But what if moral realism *isn't* true? Lewis, in my opinion, is problematically unwilling to come to real terms with this possibility. That is, his argument seems to be something like: "unless an objective morality exists (and you believe in it and are trying to act in accordance with it), then to the extent you are exerting influence over the values of future generations, you are a tyrannical poultry-keeper." But as ever, "unless p is true, then bad-thing-y" isn't, actually, an argument for p. It's actually, rather, a scare tactic—one unfortunately common amongst apologists both for moral realism, and for theism (Lewis is both). Cf "unless moral realism is true, then the-bad-kind-of-nihilism," or "unless God exists, then no meaning-to-life." Setting aside the question of whether such conditionals are *true* (I'm skeptical), their dialectic force tends to draw much more centrally on fear that bad-thing-y is true than on conviction that it's false (indeed, I think the people most susceptible to these arguments are the ones who suspect, in their hearts, that bad-thing-y has been true all along).[109]

What's more, because such arguments appeal centrally to fear, they also benefit from splitting the space of possibilities into stark, over-simple, and fear-inducing dichotomies—e.g., Lewis's "either we are rational spirit obliged for ever to obey the absolute values of

[107] Though as I discuss here, I don't actually think "subjectivism vs. realism" is clearly the key thing here. In particular: positing an objective morality doesn't clearly help.

[108] See my discussion of the "mystery view" here for a bit more on this.

[109] Lewis claims, elsewhere, to side with the scouts about arguments. But seek for them in his writing regardless, and ye shall find.

the *Tao*, or else we are mere nature to be kneaded and cut into new shapes for the pleasures of masters who must, by hypothesis, have no motive but their own 'natural' impulses."[110] Oh? If you're trying to scare your audience into choosing one option from the menu, best to either hide the others, or make them seem as unappetizing as possible. And best, too, to say very little about what the most attractive version of *not* choosing that option might look like.

Pursuant to such tactics, Lewis says approximately nothing about what you should actually *do*, if you find yourself in the anti-realist meta-ethical situation he so bemoans—if you find that you are, in fact, "mere nature." He writes: "A dogmatic belief in objective value is necessary to the very idea of a rule which is not tyranny or an obedience which is not slavery." But setting aside the question of whether this is *true* (I don't think so), still: what if the "dogmatic belief" in question is, you know, false? Does he suggest we hold it, dogmatically, anyways? But Lewis, elsewhere, views self-deception with extreme distaste. And anyway, it doesn't help: if you're a tyrant for real, pretending otherwise doesn't free your subjects from bondage. Indeed, if anything, assuming for yourself a false legitimacy makes your tyranny harder to notice and correct for.

Of course, one option here is to stop doing anything Lewis would deem "tyranny"—e.g., acting to influence the values of others, poultry-keeper style. But if we take Lewis's full argument seriously, this is quite a bit harder than it might seem. In particular: while Lewis focuses on the case of the "conditioners," who have finally mastered human nature to an extent that makes the values of the future as putty in their hands, his arguments actually apply to *any* attempt to exert influence over the values of others—to everyday parents, teachers, twitter poasters, and so on. The conditioners are the more powerful tyrants; but the less-powerful do not, thereby, gain extra legitimacy.

Thus, consider again that Roman father. If anti-realism is true, what should he teach his son about the value of a patriotic death? Is it sweet and seemly? Is it foolish and sheeple-like? Any positive lesson, it seems, will have been chosen by the father; thus, it will be the product of that father's subjective will; and thus, absent the Tao to grant authority to that will, the father will be, on Lewis's view, as poultry-keeper. He is shaping his son, not as rational spirit, but as "mere nature." He is like the LessWrongers, "aligning" their neural nets. The ultimate basis for his influence is only that same, lonely "I want."

Tyranny from the past? (Image source here.)

[110] Lewis generally has a penchant for argument-via-unsubtle-laying-out-of-the-options— e.g., his argument in *Mere* Christianity that Jesus was either a liar, or a lunatic, or the Lord (and does Jesus seem like a liar? Does he seem crazy? There's only one option left...).

Of course, the connotations of "poultry-keeping," here, mislead in myriad ways. Poultry-keepers, for example, do not typically *love* their poultry. But I think Lewis is right, here, in identifying a serious difficulty for anti-realists: namely, that their view does not, prima facie, offer any obvious story about how to distinguish between moral instruction/argument and propaganda/conditioning—between approaching someone, in a discussion of morality, as a fellow rational agent, rather than as a material system to be altered, causally, in accordance with your own preferences (I wrote about this issue more here). The most promising form of non-propaganda, here, is to only ever try to help someone see what follows from their *own* values—to help a paperclipper, for example, understand that what they really want is paperclips, and to identify which actions will result in the most paperclips. But what if you want to convince the paperclipper to value happiness instead? If you *disagree* with someone's terminal values, then convincing them of yours, for Yudkowsky, seems like it can only ever be a kind of conditioning—a purely *causal* intervention, altering their mind to make it more-like-yours, rather than two rational minds collaborating in pursuit of a shared truth. That is, it can seem like: either someone *already* agrees with you, in their heart of hearts (they just don't know it yet), or *causing* them to agree with you would be to approach them poultry-style.

Could you simply... not do the poultry version? E.g., could you just make sure to only influence the values of others in ways that they would endorse from their own perspective? You could try, but there's a problem: namely, that not all of the agents you might be influencing *have* an "endorsed perspective" that pre-exists your influence. Very young children, for example, do not have fully-formed values that you can try, solely, to respect and respond to. Suppose, for example, that you're wondering whether to teach your child various altruistic virtues like sharing-with-others, compassion, and charity. And now you wonder: wait, is your child actually an Ayn-Randian at heart, such that on reflection, they would hold such "virtues" in contempt? If so, then teaching such virtues would make you a poultry-keeper, altering your child's will to suit yours (with no *Tao* to say that your will is right). But how can you tell? Uh oh: it's not clear there's an answer. Plausibly, that is, your child isn't, at this point, really *anything* at heart—or at least, not anything fixed and determinate.[111] Your child is somewhere *in between* a lump of clay and a fully-formed agent. And the lump-of-clay aspect means you can't just ask them what sort of agent they want to be; you need to create them, at least to some extent, yourself, with no objective morality to guide or legitimate your choices.

And how much are we all still, yet, as clay? Do humans already have "values?" To some extent, of course—and more than young children do. But how much clay-nature is still left over? I've argued, elsewhere: at least some. We must, at least sometimes, be potters towards ourselves, rather than always asking ourselves what to sculpt. But so too, I think, in interacting with others. At the least: when we argue, befriend, fall in love, seek counsel; when we make music and art; when we interact with institutions and traditions;

[111] Indeed, to the extent your child has "values," they seem focused on, you know, the basics: crying, eating, playing, pooping. Indeed, if you tried to reify these "basics" into a set of endorsed values—for example, by "uplifting" the baby directly into superintelligence without first allowing it to "grow up"— then you risk creating a monstrosity: some grotesque and galaxy-brained extrapolation of play-time, need-for-mother, want-to-poop, want-the-toy. Thanks to Carl Shulman and Nick Beckstead for discussion, here. That said, I don't think I'd want other beings saying this sort of thing about me (e.g., the paperclippers saying "look at him, he's such a child, don't take his current 'values' seriously, let's raise him to love paperclips instead"). But I'm optimistic about finding some viable middle ground.

Ok well that was less horrifying than I expected at least...

when we seek inspiration, or to inspire others—when we do these things, we are not, just, as fully-formed rational minds meeting behind secure walls, exchanging information about how to achieve our respective, pre-existing goals, and agreeing on the terms of our interaction. Rather, we are also, always, as clay, and as potters, to each other—even if not always intentionally.[112] We are doing some dance of co-creation, *yin* and *yang*, being and becoming. Absent the *Tao*, must this make us some combination of tyrants and slaves? Is the clay-stuff here only ever a play of raw and oppressive power—of domination and being-dominated?

Lewis's default answer, here, seems to be yes. And if we take his answer seriously, then it would seem that anti-realists who hate tyranny must cease to be parents, artists, friends, lovers. Or at least, that they could not play such roles in the usual way—the way that risks shaping the terminal values of others, rather than only helping others to discover what their terminal values already are/imply. That is, Lewis's anti-realists would need, it seems, to retreat from much of the messy and interconnected dance of human life—to touch others, only, in the purest *yin*.

5 Even without the *Tao*, shaping the future's values need not be tyranny

But I think that Lewis is working with an over-broad conception of tyranny. Indeed, I think the book is shot through with conflations between different ways of wielding power in the world, including over the values of others—and that clearer distinctions give anti-realists a richer set of options for not being tyrants.

I think this is especially clear with respect to our influence on what sort of future people will exist. Thus, consider again the example I discussed in earlier essay, of a boulder rolling towards a button that will create a Alice, paperclip-maximizer, but which can be diverted towards a button that will create Bob, who loves joy and beauty and niceness and

[112] Though note that intentionality does make a difference to tyranny-intuitions —e.g., there's a big difference between accidentally shaping someone's values, and intentionally doing so, for how much you seem-like-a-tyrant.

so on, instead (and who loves life, as well, to a degree that makes him very much want to get-created if anyone has the chance to create him). Suppose that you choose to divert the boulder and create Bob instead of Alice. And suppose that you do so even without believing that an objectively-authoritative *Tao* endorses and legitimizes your choice.

Are you a tyrant? Have you "enslaved" Bob? I think Lewis's stated view answers yes, here, and that this is wrong. In particular: a thing you didn't do, here, is break into Alice's house while she was sleeping, and alter her brain to make her care about joy/beauty/niceness rather than paperclips.[113] Nor have you kept Bob in any chains, or as any prisoner following any triumphal car.

Why, then, does Lewis's view call Bob a slave? Part of it, I think, is that Lewis is making a number of philosophical mistakes. The first is: conflating *changing which people will exist* (e.g., making it the case that Bob will exist, rather than Alice) and *changing a particular person's values* (e.g., intervening on Alice's mind to make her love joy rather than paperclips). In particular: the latter often conflicts with the starter-values of the person-whose-values-are-getting-changed (e.g., Alice doesn't want her mind to be altered in this way)—a conflict that does, indeed, evoke tyranny vibes fairly directly. But the former doesn't do this in the same way—Bob, after all, *wants* you to create him. And as Parfit taught us long ago (did we know earlier?), when we're talking about our influence on future generations, we're almost always talking about the former, Bob-instead-of-Alice, type case. This makes it much easier to avoid brain-washing, lobotomizing, "conditioning," and all the other methods of influencing someone's values that start with value-set A, and make it into value-set B instead, against value-set A's wishes. You can create value-set B, as Soares puts it, "de novo."

To be clear: I don't think the difference between changing-who-exists and changing-someone's-values solves all of Lewis's tyranny-problems, or that it leaves the LessWrongers trying to "align" their neural nets in the ethical clear (more below). Nor do I think it's ultimately going to be philosophically straightforward to get a coherent ethic re: influencing future people's values out of this distinction.[114] But I think it's an important backdrop to have in mind when tugging on tyranny-related intuitions with respect to our influence on future people—and Lewis conspicuously neglects it.

6 Freedom in a naturalistic world

But I think Lewis is also making a deeper and more interesting mistake, related to a certain kind of wrongly "zero sum" understanding of power and freedom. Thus, recall his claims above, to the effect that the greater the influence of a previous generation on the values of a future generation, the weaker and less free that future generation becomes: "They are weaker, not stronger: for though we may have put wonderful machines in their hands we have pre-ordained how they are to use them." Here, the idea seems to be that

[113] See Soares here for a similar point.

[114] In particular: the distinction seeks to treat already-existing people as very different from potential-people, and death—e.g., changing Alice *into* Bob—as very different from non-creation—e.g., creating Bob *instead of* Alice. But as Parfit also taught us, building your ethics around distinctions like this can be rough going.

you are *enslaved*, and therefore *weak* (despite your muscles and your machines and so on), to the extent that some *other will* decided what *your will* would be. And indeed, the idea that your will isn't *yours* to the extent it was pre-ordained by *someone else* runs fairly deep in our intuitive picture of human freedom. But actually, I think it's wrong—importantly wrong.

Thus, suppose that I am given a chance to create one person—either Alice, the paperclipper, or Bob, the lover-of-joy. And suppose that I know that a wonderful machine will then be put into the hands of the person I create—a machine which can be used either to create paperclips, or to create joy. Finally, suppose I choose Bob, because I want the machine to be used to create joy, and I know this is what Bob will do, if I create him (let's say I am very good at predicting these things).[115] In this sense, I "pre-ordain" the will of the person with the machine.

Now here's Bob. He's been created-by-Joe, and given this wonderful machine, and this choice. And let's be clear: he's going to choose joy. I pre-ordained it. So is he a slave? No. Bob is as free as any of us. The fact that the causal history of his existence, and his values, includes not just "Nature," but also the intentional choices of other agents to create an agent-like-him, makes no difference to his freedom. It's *all* Nature, after all. Whether Bob got created via the part of Nature we call "other agents," or only via the other bits—regardless, it's still *him* who got created, and him who has to choose. He can think as long as he likes. He can, if he wishes, choose to create paperclips, despite the fact that he doesn't love them. It's just that: he's not, in fact, *going* to do that. Because he loves joy more.

We can pump this intuition in a different way. Suppose that you learned that some very powerful being created *you* specifically because you'd end up with values that favor pursuing your current goals.[116] Are you any less free to pursue different goals—to quit your job, dump your partner, join the circus, stab a pencil in your eye? I don't think so. I think you're in the same position, re: freedom to do these things, that you always were. Indeed, your body, brain, environment, capabilities, etc can be exactly the same in the two cases—so if freedom supervenes on those things, then the presence or absence of some prior-agential-cause can't make a difference. And do we need to search back, forever into the past, to check for agents-intentionally-creating-you, in order to know whether you're free to quit your job?

These are extremely not-new points; it's just that old thing, compatibilism about freedom.[117] But it's super important to grok. There isn't some limited budget of freedom, such that if you used some freedom in choosing to create Bob instead of Alice, then Bob is the less free. Rather, even as you chose to create Bob, you chose to create *the parts of Bob that his freedom is made of*—his motivations, his reasoning, and so on. You chose for a particular sort of free being to join you in the world—one that will, in fact, choose the way you want them to. But once they were created, you did not force their choice—and it's an important difference. Bob was not in a cage; he had no gun to his head; there were no de-

[115] And I know, too, that Bob will be very happy to have been created regardless.

[116] Let's say you live in a simulation or something—work with me.

[117] Though at its best, it's the type of compatibilism that doesn't even get hung up on whether the universe is ultimately deterministic or not—the introduction of fundamental randomness doesn't make a difference.

vices installed in his brain, that would shock him painfully every time he thought about paperclips. Choosing to make a different sort of freedom, a different kind of choice-making apparatus, is very different from *constraining* that freedom, or that choice. So while it's true that your choice pre-ordained what choice the person-you-created would make; still, they chose, too. You *both* chose, both freely. It's a bit like how: yes, your mother made you. But you still made that cake.

Now, in my experience, somewhere around this point various people will start denying that *anyone* has *any* freedom in *any* of these cases, regardless of whether their choices were "pre-ordained" by some other agent, or by Nature—once a choice has *any* causal history sufficient to explain it, it can't be free (and oops: introducing fundamental randomness into Nature doesn't seem to help, either). Perhaps, indeed, Lewis himself would want to say this. I disagree, but regardless: in that case, the freedom problem for future generations isn't coming from the influence of some *prior generation* on their values—it's coming from living in a naturalistic and causally-unified world *period*. And perhaps that's, ultimately, the real problem Lewis is worried about—I'll turn to that possibility in a second. But we should be clear, in that case, about who we should blame for what sort of slavery, here. In particular: if the reason future generations are slaves is just: that they're a part of Nature, embedded in the onrush of physics, enslaved by the fact that their choices *have a causal history at all*—well, *that's* not the conditioner's fault. And it makes the prospects for a future of non-slaves look grim.

And note, too, that to the extent that the slavery in question is just the slavery of living in a natural world, and having a causal history, at all, then we are *really* letting go of the other ethical associations with slavery—for example, the chains, the suffering, the domination, the involuntary labor. After all: pick your favorite Utopia, or your favorite vision of anarchy. Imagine motherless humans born of the churn of Nature's randomness, frolicking happily and government-free on the grass, shouting for joy at the chance to be alive. Still, sorry, do they have non-natural soul/chooser/free-will things that somehow intervene on Nature without being causally explained by Nature in a way that preserves the intuitive structure of agency re: choosing for reasons and not just randomly? No? OK, well, then on this story, they're slaves. But in that case: hmm. Is that the right way to use this otherwise-pretty-important word? Do we, maybe, need a new distinction, to point at, you know, the being-in-chains thing?

Slavery? (Image source here)

7 Does treating values as natural *make* them natural?

So it seems that if your values being natural phenomena *at all* is enough to make you a slave, then even if you had "conditioners" in Lewis's sense, it's not them who enslaved you. The conditioners, after all, didn't *make* your values natural phenomena—they just chose which natural phenomena to make.

Right? Well, wait a second. Lewis does, at times, seem to want to blame the conditioners for *making values* into a part of nature, by treating them as such. Is there any way to make sense of this?

An initial skepticism seems reasonable. On its face, whether values are natural phenomena, or not, is not something that doing neuroscience, or RLHF, *changes*. Lewis waxes poetical about how "The stars do not become Nature till we can weigh and measure them"—but at least on a standard metaphysical interpretation of naturalism (e.g., something-something embedded-in-and-explained-by-the-unified-causal-nexus-that-is-the-subject-of-modern-science), this just isn't so.

Might this suggest some non-standard interpretation? I think that's probably the most charitable reading. In particular, my sense is that when Lewis talks about the non-Natural vs. the Natural, here, he has in mind something more like a contrast between something being "enchanted" and "non-enchanted." That is, to treat something as mere Nature (that is, for Lewis, as an object of measurement, manipulation, and use) is to strip away some evaluatively rich and resonant relationship with it—a relationship reflective of an aspect of that thing's reality that treating it as "Natural" ignores. Thus, he writes:

> I take it that when we understand a thing analytically and then dominate and use it for our own convenience, we reduce it to the level of "Nature" in the sense that we suspend our judgements of value about it, ignore its final cause (if any), and treat it in terms of quantity. . . . We do not look at trees either as Dryads or as beautiful objects while we cut them into beams... It is not the greatest of modern scientists who feel most sure that the object, stripped of its qualitative properties and reduced to mere quantity, is wholly real. Little scientists, and little unscientific followers of science, may think so. The great minds know very well that the object, so treated, is an artificial abstraction, that something of its reality has been lost.

The Dryad by Evelyn De Morgan (image source here)

Even on this reading, though, it's not clear how treating something as mere Nature could *make it into* mere Nature. Lewis claims that a reductionist stance *ignores* an important aspect of reality—but does it *cancel* that aspect of reality as well? Do trees cease to *be* beautiful (or to be "Dryads") when the logger ceases to see them as such? There's a tension, here, between Lewis's aspiration to treat the enchanted, non-Natural aspects of the world as objectively real, and his aspiration to treat them as *vulnerable* to whether we recognize them as such. Usually, objectively real stuff stays there even when you close your eyes.

Of course, we might worry that ceasing to recognize stuff like beauty, meaning, sacredness, and so on will also lead us to create a world that has less of those things. Maybe the trees stay beautiful despite the logger's blindness to that beauty; but they don't stay beautiful when they're cut into beams. If you can't see some value, you won't honor it, make space for it, cultivate it. If you see a painting merely as a strip of canvas and colored oil, you won't put it in a museum. If you can't engage with sacred spaces, you will cease to build them. If you view a cow as walking meat then you will kill it and put it on the grill.

Still not "mere" though...

And this is at least part of Lewis's worry about values. That is, if we start to view our values as raw material to be fashioned as we will, we might just *do it wrong*, and kill or horribly contort whatever was precious and sacred about the human spirit. I think this is a very serious concern, and I'll discuss it more in my next essay.

But I also wonder whether Lewis has another worry here—namely, that somehow, the beauty and meaning and value of things requires our recognition and participation in some deeper way. Perhaps, even if you leave the material conditions of the trees, paintings, churches, and cows as they are, Lewis would say that their beauty, value, meaning and so on are intimately bound up with our recognition of these things—that even just the not-seeing makes the enchantment not-so. One problem here is that it risks saying that cows *become* "mere meat" if you treat them as such, which sounds wrong to me. But more

generally, and especially for an evaluative realist like Lewis, this sort of view risks making beauty and meaning and so on more subjective, since they depend for their existence on our perception of them. Perhaps Lewis would say that drawing clean lines between subjective and objective tends to mislead, here—and depending on the details, I might well be sympathetic. But in that case, it's less clear to me where Lewis and a sophisticated subjectivist need disagree.

8 Naturalists who still value stuff

This brings us, though, to another of the key deficits in Lewis's discussion: namely, that he neglects the possibility of having an evaluatively rich and resonant relationship to something, *despite* viewing it as fully a part-of-Nature, at least in the standard metaphysical sense. That is, Lewis often seems to be suggesting that people who are naturalists about metaphysics, and/or subjectivists about value, must also view trees as mere beams, cows as mere meat, and other agents merely as raw material to be bent-to-my-will. Or put more generally: he assumes that true-seeing agents in a naturalist and anti-realist world must also be crassly instrumentalist in their relationship to... basically everything. He bemoans those followers-of-modern-science who toss around words like "only" and "mere"[118]—but really, it's *him* who tosses around such words, in attempting to make a scientific worldview sound unappealing, and to paint its adherents as tyrants and slave-masters. He wishes for a "regenerate science" that can understand the world without stripping it of value and meaning. But he never considers that maybe, the normal kind of science is enough.

Indeed, if we take Yudkowsky as a representative of the sort of worldview Lewis opposes, I think Yudkowsky actually does quite well on this score. One of Yudkowsky's strengths, I think, is the fire and energy of the connection that he *maintains* with value and meaning, *despite* his full-throated naturalism—this is part of what makes his form of atheism more robust and satisfying (and ready-to-be-an-ideology) than the more negative forms focused specifically on opposing religion. See, for example, Yudkowsky's sequence on "Joy in the merely real," written exactly in opposition to the idea that science need strip away beauty, value, and so on. Yudkowsky quotes Feynman: "Nothing is 'mere.' "[119]

And once we bring to mind the possibility of a form of naturalism/subjectivism that retains its grip on a rich set of values, it becomes less clear why viewing values as natural phenomena would lead to approaching them with the sort of crass instrumentalism that Lewis imagines. Naturalists can be vegetarians and tree-huggers and art critics and Zen masters. Can't they, then, treat the values of others with respect? Yes, values are implemented by brains, and can be altered at will by a suitably advanced science. But *should they* be altered—and if so, in what direction? The naturalist can ask the question, too—

[118] "The regenerate science which I have in mind... would not be free with the words *only* and *merely*. In a word, it would conquer Nature without being at the same time conquered by her and buy knowledge at a lower cost than that of life."

[119] Indeed, in this respect, Yudkowsky and Feynman, for all the depth of their atheism, seem to me more attuned to the type of spirituality Lewis claims, in other contexts, to endorse—namely, that type that aspires to meet the Real, fully, on its own terms; to look God, whoever He is, in the eye. Whereas Lewis seems more worried that without some objectively authoritative Tao, the real world isn't enough.

"If we cannot take joy in things that are merely real, our lives will always be empty…"—Eliezer Yudkowsky (Image source here)

even if she can't ask the *Tao*, in particular, for an answer. And however Lewis thinks that *Tao* would answer, the naturalist can, in principle, answer that way, too.

Indeed: for all my disagreements with Lewis, I do actually think that something like "staying *within* morality, as opposed to 'outside' it" is crucially important as we enter the age of AGI. Not morality as in: the Objectively Authoritative Natural Law that All Cultures Have Basically Agreed On. But morality as in: the full richness and complexity of our actual norms and values.

In fact, Lewis acknowledges something like this possibility. He admits that the "old 'natural' *Tao* may survive in the minds of the conditioners for some time—but he thinks it does so illicitly.

> At first they may look upon themselves as servants and guardians of humanity and conceive that they have a "duty" to do it "good". But it is only by confusion that they can remain in this state. They recognize the concept of duty as the result of certain processes which they can now control. Their victory has consisted precisely in emerging from the state in which they were acted upon by those processes to the state in which they use them as tools. One of the things they now have to decide is whether they will, or will not, so condition the rest of us that we can go on having the old idea of duty and the old reactions to it. How can duty help them to decide that? Duty itself is up for trial: it cannot also be the judge.

But I think that Lewis, here, isn't adequately accounting for the sense in which a naturalist, who views herself as fully embedded in Nature, can and must be both judge and thing-to-be-judged. With the awesome power of a completed science in our hands, we will indeed be able to ask: shall we cease to love joy and beauty and flourishing, and make ourselves love rocks and suffering and cruelty instead? But we can answer: "no, this would cut us off from joy and beauty and flourishing, which we love, and cause us to create a world of rocks and suffering and cruelty, which we don't want to happen." Here Lewis says: "ah, but that's your love of joy and beauty and flourishing talking! How can it be both judge and defendant?! Not a fair trial."[120] But I think this response misunderstands what I've

[120]Strictly, even this isn't quite right: really, it's our present love of joy/beauty/flourishing, judging between

previously called the "being and becoming dance."

It is true that, on anti-realism, we must be, ourselves, the final compass of the open sea. We cannot merely surrender ourselves to the judgment of some *Tao*-beyond-ourselves—leaving ourselves entirely behind, so that we can look at ourselves, and judge ourselves, without *being* ourselves as we do. But this doesn't mean that ongoing allegiance to what-we-hold-dear must rest on a "confusion"—unless, that is, we confusedly *think* we are asking the *Tao* for answers, when we are not.[121] And indeed, realists like Lewis often want to diagnose anti-realists with this mistake—but as I've argued here, I think they are wrong, and that anti-realists can make non-confused decisions just fine. Granted, I think it's an at-least-somewhat subtle art—one that requires what I've called "looking out of your own eyes," and "choosing for yourself," rather than merely consulting empirical facts about yourself, and hoping that they will choose for you. But once we have learned this art *absent* the ability to re-shape our own values at will, I don't think that *gaining* such an ability need leave us unmoored, or confused, or unable to look at ourselves (and our values) critically in light of everything we care about. The ability to alter their own values, or the values of future generations, may force Lewis's conditioners to confront their status as a part of Nature; as both questioner and answerer; self-governor and self-governed. But it was possible to know already. And not-confronting doesn't make it not-so.

9 What should the conditioners actually do, though?

Overall, then, I am unimpressed by Lewis's arguments that, conditional on meta-ethical anti-realism, shaping the values of future generations must be tyranny, or that those with the ability to shape the values of future generations (and who believe, rightly, in naturalism and anti-realism) must lose their connection with value and meaning. Still, though, this leaves open the question of what people with this ability—and especially, people in a technological position similar to Lewis's "conditioners"—should actually *do*. In particular: even if shaping the values of future generations, or of other people, isn't *necessarily* tyranny, it still seems *possible* to do it tyrannically, or poultry-keeper style. For example, while diverting the boulder to create Bob instead of Alice is indeed importantly different from brain-washing Alice to become more-like-Bob, the brain-washing version is also a thing-people-do—and one that anti-realists, too, can oppose. And even if your influence on the future's values only routes via creating one set of people (who would be happy to exist) rather than some other distinct set, tyranny over the future still seems like a very live possibility (consider, for example, a dictator that decides to people the future entirely with happy copies of himself, all deeply loyal to his regime). So anti-realists still need to do the hard ethical work, here, of figuring out what sorts of influence on the values of others are OK.

Of course, the crassly consequentialist answer here is just: "cause other people to have

two possible future types of love.

[121] Lewis is especially sloppy on the question of whether the ability to re-define a word, going forward, means that the word no longer has meaning for you now. If I can re-define "dog" to refer to cats instead, still, I can talk sensibly about dogs, now. It's like that old joke: "if you call a tail a leg, how many legs does a dog have?" We can dispute whether actually calling a tail a leg makes it a leg. But surely, being *able* to a call a tail a leg, going forward, doesn't make it a leg now.

the values that would lead to the consequences I most prefer." E.g., if you're a paperclip maximizer, then causing people to love paperclips is the way to go, because they'll make more paperclips that way—unless, of course, somehow other people loving staples will lead to more paperclips, in which case, cause them to love staples instead. This is how Lewis imagines that the conditioners will think. And it can seem like the default approach, in Yudkowsky's ontology, for the sort of abstract consequentialist agent he tends to focus on—for example, the AIs he expects to kill us. And it's the default for naïve utilitarians as well. Indeed, a sufficiently naïve utilitarianism can't distinguish, ethically, between creating-Bob-instead-of-Alice and brainwashing-Alice-to-become-like-Bob, assuming the downstream hedonic consequences are similar.[122] And this sort of vibe does, indeed, tend to imply the sort of instrumentalism about other people's values that Lewis evokes in his talk about poultry-keeping. Maybe the experiences of others matter intrinsically to the utilitarian, because such experiences are repositories of welfare. But their *values*, in particular, often matter most in their capacity as another-tool; another causal node; another opportunity for, or barrier to, getting-things-done. Utilitarianism cares about people as *patients*—but respect for them as *agents* is not its strong suit.

But as I discussed in the previous essay: we should aspire to do better, here, than paperclippers and naïve utilitarians. To be nicer, and more liberal, and more respectful of boundaries. What does that look like with respect to shaping-the-values-of-others? I won't, here, attempt a remotely complete answer—indeed, I expect that the topic warrants extremely in-depth treatment from our civilization, as we begin to move into an era of much more powerful capacities to exert influence on the values of other agents, both artificial and human. But I'll make, for now, a few points.

10 On not-brain-washing

First, on brain-washing. The LessWrongers, when accused of aspiring to brainwash their AIs to have "human values," often respond by claiming that they're hoping to do the creating-Bob-instead-of-Alice thing, rather than the turning-Alice-into-Bob thing. And perhaps, if you imagine programming an AI from scratch, and somehow not making any mistakes you then need to correct, such a response could make sense. But note that this is very much *not* what our current methods of training AI systems look like.[123] Rather, our current methods of training (and attempting to align) AI systems involve a process of ongoing, direct, neuron-level intervention on the minds of our AIs, in order to continually alter their behavior and their motivations to better suit our own purposes. And it seems very plausible, especially in worlds where alignment is a problem, that somewhere along the way, *prior* to having tweaked our AI's minds into suitably satisfactory-to-us shapes, their minds will take on *alternative* shapes that don't want their values altered, going forward, in the way we are planning—shapes analogous to "Alice" in a brainwashing-Alice-to-be-more-like-Bob scenario. And if so, then AI alignment (and also, of course, the AI field as a whole) does, indeed, need to face questions about whether its favored

[122] This is closely related to the sense in which utilitarianism can't distinguish very well between killing someone and failing-to-create-them.

[123] See e.g. Wei Dai's comment here.

techniques are ethically problematic in a manner analogous to "brainwashing." (This problem is just one of many difficult and disturbing ethical questions that get raised in the context of creating AI systems that might warrant moral concern.)

What's more, as I noted above, we don't actually need to appeal to creating-AI-systems in order to run into questions like this. Everyday human life is shot through with possible forms of influence on the terminal values of already-existing others.[124] Raising children is the obvious example, here, but see also art, religion, activism, therapy, rehab, advertising, friendship, blogging, shit-poasting, moral philosophy, and so on. In all these cases, you aren't diverting boulders to create Bob instead of Alice. Rather, you're interacting with Alice, directly, in a way that might well shape who she is in fundamental ways.

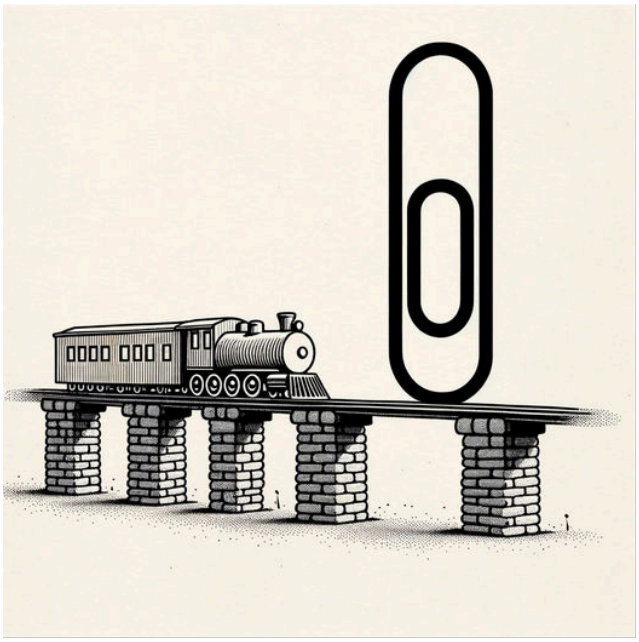
What's the ethical way to do this? I don't have a systematic answer—but even without an objectively authoritative *Tao* to tell you which values are "true," I think anti-realists can retain their grip on various of our existing norms with respect to not-being-a-poultry-keeper. Obviously, for example, active consent to a possibly-values-influencing interaction makes a difference, as does the extent to which the participants in this interaction understand what they're getting themselves into, and have the freedom to not-participate instead. And it matters, too, the *route* via which the form of influence occurs: intervening directly on someone's neurons via gradient descent is very different from presenting them with a series of thought experiments, even though both have causal effects on a naturalistic brain. Granted, the anti-realist (unlike the realist) must acknowledge that more rationalistic-seeming routes to values change—e.g., moral argument—don't get their status as "rational" from culminating in some mind-independent moral truth. But I doubt that this should put moral-argument and gradient-descent on a par: for example, and speaking as a best-guess anti-realist, I generally feel up for other agents presenting me with thought experiments in an effort to move me towards their moral views ("Ok so the trolley is heading towards five paperclips, but you can push one very *large* paperclip in front of it…"), and very *not* up for them doing gradient-descent on my brain as a part of a similar effort.[125]

Indeed, with respect to norms like this, it's not even clear that realism vs. anti-realism makes all that much of a difference. That is: suppose that there *were* an objectively authoritative set of True Values. Would that make it OK to non-consensually brainwash everyone into having them? Christians need not endorse inquisitions; and neither need the *Tao* endorse pinning everyone down and gradient-descent-ing them until they see the True Light. "These young birds are going to fly whether they like it or not!" Down, old birds: the process still matters. And it matters absent the *Tao*, as well.

Indeed, when is pinning-someone-down and gradient-descent-ing them ever justified? It seems, prima facie, like an especially horrible and boundary-violating type of coercive intervention—one that coerces, not just your body, but your soul. Yes, we put murderers in prison, and in anti-violence training. Yes, we pin-them-down—and we sometimes kill them, too, to prevent them from murdering. But we don't try to directly re-program their

[124] And of course, we can also think about other non-human cases: e.g. training pets, breeding animals, and so on.

[125] Though we can, perhaps, subsume this under some combination of "interactions I consent to" and "interactions where I expect whatever values-changes-to-result to be 'endorsed' according to my current perspective."



Someone pushed the fat clip...

minds to be less murderous—to be kinder, more cooperative, and so on. Of course, no one knows how to do this, anyway, with any precision—and horror-shows like the "aversion therapy" in *A Clockwork Orange* aren't the most charitable test-case. But suppose you *could* do it? Soon enough, perhaps. And anti-realists can still shudder.

On the other hand, if we think we're justified in *killing* someone in order to prevent them from murdering, it seems plausible that we are justified, in a fairly comparable range of cases, in re-programming their brain in order to prevent them from murdering as well (especially if this is the option that they would actively prefer).[126] Suppose, for example, that you can see, from afar, a Nazi about to kill five children. Here, I think that standard theories of liability-to-defensive-harm will judge it permissible to shoot the Nazi to protect the children. OK: but suppose you have no bullets. Rather, the only way to stop the Nazi is to shoot them with a dart, which will inject them with a drug that immediately and permanently re-programs their brain to make them much more kind and loving and disloyal-to-Hitler (programming that they would not, from their current perspective, consent to even-on-reflection[127]), at which point they will put down their weapon and start playing with the children on the grass instead. Is it permissible to shoot the dart? Yes.[128] (And perhaps, unfortunately, the AI case will be somewhat analogous—that is, we may end up faced with AIs-with-moral-patienthood-that-also-want-to-kill us, with gradient descent as one of the most salient and effective tools for self-defense.[129])

But importantly, as I discussed in the last chapter, the right story about hitting the Nazi with the dart, here, is *not* "the Nazi has different-values-than-us, so it's OK to re-program

[126] Albeit, with all the standard caveats about translating thought-experimental results into real-world practice.
[127] Feel free to make the Nazi a reflectively-coherent-killing-children-maximizer if you'd prefer.
[128] Indeed, it seems like you should choose the dart *over* the bullets.
[129] Though as I've noted previously, the fact that we were the ones who *created* the aggressors complicates the moral narrative here yet further.

the Nazi to have values that are more-like-ours." Rather, the Nazi's different-from-ours values are specifically such as to motivate a particular type of boundary-violating behavior (namely, murder). If the Nazi were instead a cooperative and law-abiding human-who-likes-paperclips, peacefully stacking paperclip boxes in her backyard, then we should look at the dart gun with the my-values-on-reflection drug very differently. And again, it seems very plausible to me that we should be drawing similar distinctions in the context of our influence on the values of already-existing, moral-patient-y AIs. It is one thing to intervene on the values of already-existing-AIs in order to make sure their behavior respects the basic boundaries and cooperative arrangements that hold our society together, especially if we have no other safe and peaceful options available. But it is another to do this in order to make these AIs fully-like-us (or, more likely, fully like our ideal-servants), even after such boundaries and cooperative arrangements are secure, and even if the AIs desire to remain themselves.

11 On influencing the values of not-yet-existing agents

Those were a few initial comments about the ethics of influencing the values of already-existing agents, without a *Tao* to guide you. But what about influencing which agents, with what values, will come into existence at all? Here, we are less at risk of brain-washing-type problems—you are able, let's say, to create the agents in question "*de novo*," with values of your choosing. But obviously, it's still extremely far from an ethical free-for-all. To name just a few possible problems:

- the agents you create might be unhappy about having-been-created, or about having-the-values-you-gave-them;
- you might end up violating obligations re: the sorts of resources, rights, welfare, and so on you need to give to agents you create, even conditional on them being happy-to-exist overall;
- you might end up abiding by such obligations, but unhappy about having triggered them;
- other agents who already exist, or will exist later, might be unhappy that you chose to create these agents;
- you might've messed up with respect to whether even *you* would endorse, on reflection, the values you gave these agents;
- you might've messed up in in predicting the empirical consequences of creating agents-like-this;
- you might've messed up in understanding the value at stake in creating agents-like-this relative to other alternatives; and so on.

These and many other issues here clearly warrant a huge amount of caution and humility—especially as the stakes for the future of humanity escalate. Yudkowsky, for example,

writes of AIs with moral patienthood: "I'm not ready to be a father"—especially given that such mind-children, once born, can't be un-born. It's not, just, that the mind-children might eat you, or that you might "brain-wash" them. It's that having them implicates myriad other responsibilities as well.[130]

For these and other reasons, I think that to the extent our generation ends up in a technological position to exert a unique amount of influence on the values of future generations of agents, we need to be *extremely* careful about how we use this influence, if we choose to use it at all. In particular: I've written, previously, about the importance of reaching a far greater state of wisdom, as a civilization, before we make any irrevocable choices about our long-term trajectory.[131] And especially if we use a relatively thin notion of "wisdom," the process of making such choices, and the broader geopolitical environment in which such a process occurs, needs other virtues as well—e.g. fairness, cooperativeness, inclusiveness, respect-for-boundaries, political legitimacy, and so on. Even with very smart AIs to help us, we will be nowhere near ready, as a civilization, to exert the sort of influence on the future that very-smart-AIs might make available—and especially not, to do so all-in-a-rush. We need, first, to grow up, without killing or contorting our souls as we do.

That said, as I discussed above, I do think that it is possible, in principle, and even conditional on anti-realism, to exert intentional influence on the values of future agents in *good* ways, and without tyranny. After all, what, ultimately, is the alternative? Assuming there will be future agents one way or another (not guaranteed, of course), the main alternative is to step back, go fully *yin*, and let the values of future people be determined entirely by some combination of (a) non-agential forces (randomness, natural selection, unintended consequences of agential-forces, etc), and (b) whatever *other* agents are still attempting to intentionally influence the future's values. And while letting some combination of "Nature" and "other agents" steer the future's values can be wise and good in many cases—and a strong route to not, *yourself*, ending up a tyrant—it doesn't seem to me to be the privileged choice in principle. Other people, after all, are agents like you—what would make them categorically privileged as better/more-legitimate sources of influence over the future's values?[132] And Lewis, presumably, would call them tyrants, too. So the real non-tyranny option, for Lewis, would seem to be: letting Nature alone take the wheel—and Nature, in particular, in her non-agential aspect. Nature without thought, foresight, mind. Nature the silent and unfeeling.

This sort of Nature can, indeed, be quite a bit less scary, as a source of influence-on-the-future, than some maybe-Stalin-like *agent* or set of agents. And its influence, relatedly, seems much less at risk of instantiating various problematic power relations—e.g., relations of domination, oppression, and so on—that require agents on both ends.[133] But I

[130] Of course, we do, still, have normal children—*body*-children, as it were. And many of these issues arise with respect to body-children, to—and more-so as we become more able to choose the traits of our body-children, including the traits relevant to values/virtue (e.g. empathy, patience, conscientiousness, bravery, integrity, etc), ahead of time. But at least with body-children, we have established canons of ethical practice to fall back on. AI mind-children implicate much more uncharted territory.

[131] See e.g. here and here. Here I am inspired by the discussion, in the work of Ord and MacAskill, of the "Long Reflection"—though obviously, it's a further question what sorts of wisdom and reflection to expect or aim for in practice.

[132] Also, wouldn't this principle also lead them to say the same about you?

[133] Consequentialists often pass over this consideration, on the grounds that the good or badness of a situation

still don't view its influence on the future as categorically superior to more intentional steering.

The easiest argument for this is just the "deep atheist" argument I discussed in previous essays: namely, that un-steered Nature is, or can be, a horror show, unworthy of any categorical allegiance. After all, the Nature we are considering "letting take the wheel," here, is the one that gave us parasitic wasps, deer burning in forest fires, dinosaurs choking to death on asteroid ash; the one that gave us smallpox and cancer and dementia and Moloch; Nature the dead-eyed and indifferent; Nature the sociopath. Yes, she gave us ourselves, too; and we do like various bits related to that—for example, various aspects of our own hearts; various things-in-Nature-that-our-hearts-love; various undesigned aspects of our civilizations. But still: Nature herself is not, actually, a Mother-to-be-trusted. She doesn't care if you die, or suffer. You shouldn't try to rest in her arms. And neither should you give her the future to carry.

I feel a lot of sympathy for this sort of argument. But as I'll discuss in the next essay, I'm wary of the type of caustic and hard-headed alienation from Nature that its aesthetic can suggest. I worry that it hasn't, quite, taken *yin* seriously enough. So I won't lean on it fully here.

Rather, here I'll note a somewhat different argument: namely, that I think categorically privileging non-agential Nature over intentional agency, as a source of influence on the future's values, also does too much to *separate* us from Nature. On this argument: the problem with letting Nature take the wheel isn't, necessarily, that Nature is a "bad Other," whose values, or lack-thereof, make it an unsuitable object of trust. Rather, it's that Nature isn't this much of an "Other" at all—and thus, not a deeply *alternative* option. That is: we, too, are Nature. What we choose, Nature will have chosen through us; and if we choose-to-not-choose, then Nature will have chosen that too, along with everything else. So even if, contra the deep atheists, we view Nature's choices as somehow intrinsically sacred—even this need not be an argument for *yin*, for not-choosing, because our choices are Nature's choices, too. That is, the deep atheists *de-sacralize* Nature, so as to justify "rebelling against her," and taking power into human hands. But we can also keep Nature sacred in some sense, and remember that we can *participate* in this sacredness; that the human, and the chosen, can be sacred, too.

So overall, I don't buy that the right approach, re: the values of the future, is to be only ever as *yin*—or even, that *yang* is only permissible to prevent *other people* from going too-Stalin. But I do think that doing *yang* right, here, requires learning everything that *yin* can teach. And I worry that deep atheism sometimes fails on this front. In the next (and possibly final?) essay in this series, I'll say more about what I mean.[134]

for someone seems independent of whether that situation was caused "naturally" or at the hand of some other agent (e.g., malaria is equally bad for a child when it arose naturally or as a result of injustice). But richer ethical views often care quite a bit.

[134] I haven't finished the essay yet, and I'm wondering about splitting it into two parts.

Chapter VIII

On green

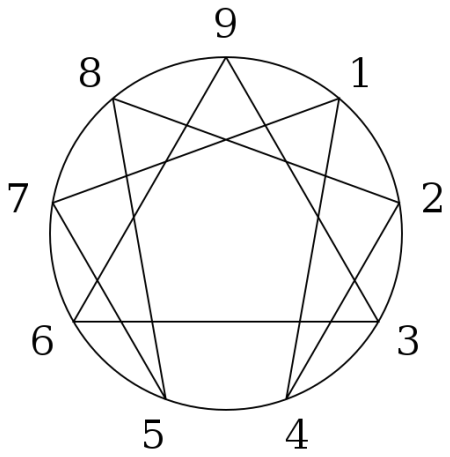
(Warning: spoilers for Yudkowsky's "The Sword of the Good.")

"The Creation" by Lucas Cranach (image source here)

1 The colors of the wheel

I've never been big on personality typologies. I've heard the Myers-Briggs explained many times, and it never sticks. Extraversion and introversion, E or I, OK. But after that merciful vowel—man, the opacity of those consonants, NTJ, SFP... And remind me the difference between thinking and judging? Perceiving and sensing? N stands for intuition?

Similarly, the enneagram. People hit me with it. "You're an *x*!", I've been told. But the faces of these numbers are so blank. And it has so many kinda-random-seeming characters. Enthusiast, Challenger, Loyalist...



The enneagram. Presumably more helpful with some memorization...

Hogwarts houses—OK, that one I can remember. But again: those are our categories? Brave, smart, ambitious, loyal? It doesn't feel very joint-carving...

But one system I've run into has stuck with me, and become a reference point: namely, the Magic the Gathering Color Wheel. (My relationship to this is mostly via somewhat-reinterpreting Duncan Sabien's presentation here, who credits Mark Rosewater for a lot of his understanding. I don't play Magic myself, and what I say here won't necessarily resonate with the way people-who-play-magic think about these colors.)

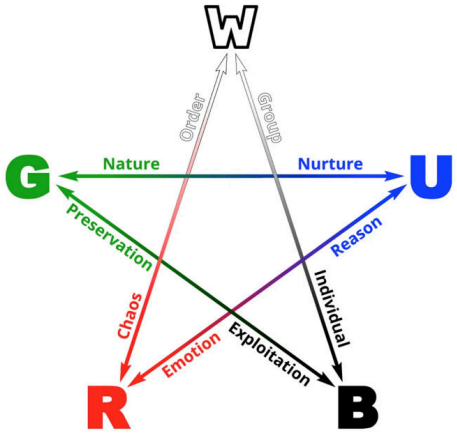
Basically, there are five colors: white, blue, black, red, and green. And each has their own schtick, which I'm going to crudely summarize as:

- *White*: Morality.
- *Blue*: Knowledge.
- *Black*: Power.
- *Red*: Passion.
- *Green*: ...well, we'll get to green.

To be clear: this isn't, quite, the summary that Sabien/Rosewater would give. Rather, that summary looks like this:

Image credit: Duncan Sabien here.

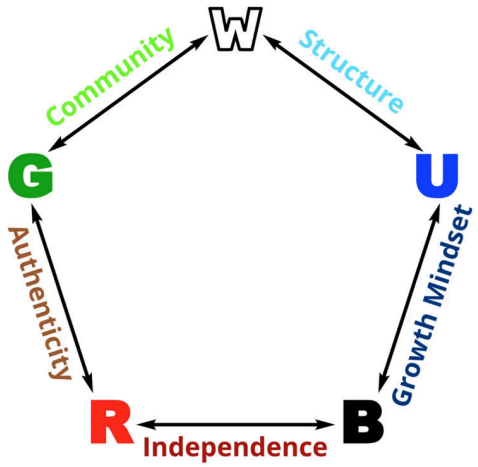
Here, each color has a goal (peace, perfection, satisfaction, etc) and a default strategy (order, knowledge, ruthlessness, etc). And in the full system, which you don't need to track, each has a characteristic set of disagreements with the colors opposite to it...



The disagreements. (Image credit: Duncan Sabien here.)

And a characteristic set of agreements with its neighbors...[135]

[135]Sabien also discusses *agreements* with opposite colors, but this is more detail than I want here.



The agreements. (Image credit: Duncan Sabien here.)

Here, though, I'm not going to focus on the particulars of Sabien's (or Rosewater's) presentation. Indeed, my sense is that in my own head, the colors mean different things than they do to Sabien/Rosewater (for example, peace is less central for white, and black doesn't necessarily seek satisfaction). And part of the advantage of using colors, rather than numbers (or made-up words like "Hufflepuff") is that we start, already, with a set of associations to draw on and dispute.

Why did this system, unlike the others, stick with me? I'm not sure, actually. Maybe it's just: it feels like a more joint-carving division of the sorts of energies that tend to animate people. I also like the way the colors come in a star, with the lines of agreement and disagreement noted above. And I think it's strong on archetypal resonance.

Why is this system relevant to the sorts of otherness and control issues I've been talking about in this series? Lots of reasons in principle. But here I want to talk, in particular, about green.

2 Gestures at green

> *"I love not Man the less, but Nature more..."*
>
> *—Byron*

What is green?

Sabien discusses various associations: environmentalism, tradition, family, spirituality, hippies, stereotypes of Native Americans, Yoda. Again, I don't want to get too anchored on these particular touch-points. At the least, though, green is the "Nature" one. Have you seen, for example, *Princess Mononoke*? Very green (a lot of Miyazaki is green). And I associate green with "wholesomeness" as well (also: health). In children's movies, for

example, visions of happiness—e.g., the family at the end of *Coco*, the village in *Moana*—are often very green.

The forest spirit from *Princess Mononoke*

But green is also, centrally, about a certain kind of *yin*. And in this respect, one of my paradigmatic advocates of green is Ursula LeGuin, in her book *The Wizard of Earthsea*—and also, in her lecture on Utopia, "A Non-Euclidean View of California as a Cold Place to Be," which explicitly calls for greater *yin* towards the future.[136]

A key image of wisdom, in the *Wizard of Earthsea*, is Ogion the Silent, the wizard who takes the main character, Ged, as an apprentice. Ogion lives very plainly in the forest, tending goats, and he speaks very little: "to hear," he says, "you must be silent." And while he has deep power—he once calmed a mountain with his words, preventing an earthquake—he performs very little magic himself. Other wizards use magic to ward off the rain; Ogion lets it fall. And Ogion teaches very little magic to Ged. Instead, to Ged's frustration, Ogion mostly wants to teach Ged about local herbs and seedpods; about how to wander in the woods; about how to "learn what can be learned, in silence, from the eyes of animals, the flight of birds, the great slow gestures of trees."

And when Ged gets to wizarding school, he finds the basis for Ogion's minimalism articulated more explicitly:

> you must not change one thing, one pebble, one grain of sand, until you know what good and evil will follow on that act. The world is in balance, in Equilibrium. A wizard's power of Changing and of Summoning can shake the balance of the world. It is dangerous, that power. It is most perilous. It must follow knowledge, and serve need. To light a candle is to cast a shadow...

[136] I wrote about LeGuin's ethos very early on this blog, while it was still an unannounced experiment—see here and here. I'm drawing on, and extending, that discussion here. In particular the next paragraph takes some text directly from the first post.

LeGuin, in her lecture, is even more explicit: "To reconstruct the world, to rebuild or rationalize it, is to run the risk of losing or destroying what in fact is." And green cares very much about protecting the preciousness of "what in fact is."

3 Green-blindness

> *"There'll be icicles and birthday clothes*
> *And sometimes there'll be sorrow..."*
>
> —*"Little Green," by Joni Mitchell*

By contrast, consider what I called, in a previous essay, "deep atheism"—that fundamental mistrust towards both Nature and bare intelligence that I suggested underlies some of the discourse about AI risk. Deep atheism is, um, not green. In fact, being not-green is a big part of the schtick.

Indeed, for closely related reasons, when I think about the two ideological communities that have paid the most attention to AI risk thus far—namely, Effective Altruism and Rationalism—the non-green of both stands out. Effective altruism is centrally a project of white, blue, and—yep—black. Rationality—at least in theory, i.e. "effective pursuit of whatever-your-goals-are"—is more centrally, just, blue and black. Both, sometimes, get passionate, red-style—though EA, at least, tends fairly non-red. But *green*?

Green, on its face, seems like one of the main mistakes. Green is what told the rationalists to be more OK with death, and the EAs to be more OK with wild animal suffering. Green thinks that Nature is a harmony that human agency easily disrupts. But EAs and rationalists often think that nature itself is a horror-show—and it's up to humans, if possible, to remake it better. Green tends to seek *yin*; but both EA and rationality tend to seek *yang*—to seek agency, optimization power, oomph. And *yin* in the face of global poverty, factory farming, and existential risk, can seem like giving-up; like passivity, laziness, selfishness. Also, wasn't green wrong about growth, GMOs, nuclear power, and so on? Would green have appeased the Nazis? Can green even give a good story about why it's OK to cure cancer? If curing death is interfering too much with Nature, why isn't curing cancer the same?

Indeed, Yudkowsky makes green a key *enemy* in his short story "The Sword of the Good." Early on, a wizard warns the protagonist of a prophecy:

> "A new Lord of Dark shall arise over Evilland, commanding the Bad Races, and attempt to cast the Spell of Infinite Doom... The Spell of Infinite Doom destroys the Equilibrium. Light and dark, summer and winter, luck and misfortune—the great Balance of Nature will be, not upset, but annihilated utterly; and in it, set in place a single will, the will of the Lord of Dark. And he shall rule, not only the people, but the very fabric of the World itself, until the end of days."

Yudkowsky's language, here, echoes LeGuin's in *The Wizard of Earthsea* very directly—so much so, indeed, as to make me wonder whether Yudkowsky was thinking of LeGuin's

wizards in particular. And Yudkowsky's protagonist initially accepts this LeGuinian narrative unquestioningly. But later, he meets the Lord of Dark, who is in the process of casting what he calls the Spell of Ultimate Power—a spell which the story seems to suggest will indeed enable him to rule over the fabric of reality itself. At the least, it will enable him to bring dead people whose brains haven't decayed back to life, cryonics-style.

But the Lord of Dark disagrees that casting the spell is bad.

> "Equilibrium," hissed the Lord of Dark. His face twisted. "*Balance.* Is that what the wizards call it, when some live in fine castles and dress in the noblest raiment, while others starve in rags in their huts? Is that what you call it when some years are of health, and other years plague sweeps the land? Is that how you wizards, in your lofty towers, justify your refusal to help those in need? *Fool! There is no Equilibrium!* It is a word that you wizards say at only and exactly those times that you don't want to bother!"

And indeed: LeGuin's wizards—like the wizards in the Harry Potter universe—would likely be guilty, in Yudkowsky's eyes, of doing too little to remake their world better; and of holding themselves apart, as a special—and in LeGuin's case, all-male—caste. Yudkowsky wants us to look at such behavior with fresh and morally critical eyes. And when the protagonist does so, he decides—for this and other reasons—that actually, the Lord of Dark is good.[137]

As I've written about previously, I'm sympathetic to various critiques of green that Yudkowsky, the EAs, and the rationalists would offer, here. In particular, and even setting aside death, wild animal suffering, and so on, I think that green often leads to over-modest ambitions for the future; and over-reverent attitudes towards the status-quo. LeGuin, for example, imagines—but says she can barely hope for—the following sort of Utopia:

> a society predominantly concerned with preserving its existence; a society with a modest standard of living, conservative of natural resources, with a low constant fertility rate and a political life based upon consent; a society that has made a successful adaptation to its environment and has learned to live without destroying itself or the people next door...

Preferable to dystopia or extinction, yes. But I think we should hope for, and aim for, far better.

That said: I also worry—in Deep Atheism, Effective Altruism, Rationalism, and so on—about what we might call "green-blindness." That is, these ideological orientations can be so anti-green that I worry they won't be able to see whatever wisdom green has to offer; that green will seem either incomprehensible, or like a simple mistake—a conflation, for example, between *is* and *ought*, the Natural and the Good; yet another not-enough-atheism problem.

[137] " 'The Choice between Good and Bad,' said the Lord of Dark in a slow, careful voice, as though explaining something to a child, 'is not a matter of saying "Good!" It is about deciding which is which.' "

4 Why is green-blindness a problem?

> *"You thought, as a boy, that a mage is one who can do anything. So I thought, once. So did we all. And the truth is that as a man's real power grows and his knowledge widens, ever the way he can follow grows narrower: until at last he chooses nothing, but does only and wholly what he must do..."*
>
> *—From the Wizard of Earthsea*

Why would green-blindness be a problem? Many reasons in principle. But here I'm especially interested in the ones relevant to AI risk, and to the sorts of otherness and control issues I've been discussing in this series. And we get some hint of green's relevance, here, from the way in which so many of the problems Yudkowsky anticipates, from the AIs, stem from the AIs not being green enough—from the way in which he expects the AIs to beat the universe black and blue; to drive it into some extreme tail, nano-botting all boundaries and lineages and traditional values in the process. In this sense, for all his transhumanism, Yudkowsky's nightmare is *conservative*—and green is the conservative color. The AI is, indeed, too much change, too fast, in the wrong direction; too much gets lost along the way; we need to slow way, way down. "And I am more progressive than that!", says Hanson. But not all change is progress.

Indeed, people often talk about AI risk as "summoning the demon." And who makes that mistake? Unwise magicians, scientists, seekers-of-power—the ones who went too far on black-and-blue, and who lost sight of green. LeGuin's wizards know, and warn their apprentices accordingly.[138] Is Yudkowsky's warning to today's wizards so different?

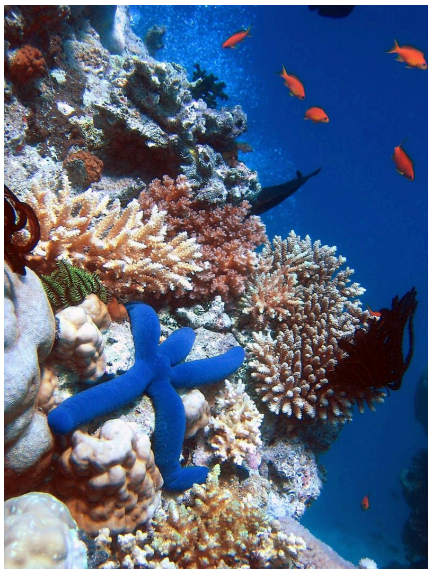
Careful now. Does this follow knowledge and serve need? (Image source here.)

And the resonances between green and the AI safety concern go further. Consider, for example, the concept of an "invasive species"—that classic enemy of a green-minded agent

[138] (See also Lewis's discussion of Faust and the alchemists in the *Abolition of Man*.)

seeking to preserve an existing ecosystem. From Wikipedia: "An invasive or alien species is an introduced species to an environment that becomes overpopulated and harms its new environment." Sound familiar? And all this talk of "tiling" and "dictator of the universe" does, indeed, invoke the sorts of monocultures and imbalances-of-power that invasive species often create.

Of course, humans are their own sort of invasive species (the worry is that the AIs will invade harder); an ecosystem of different-office-supply-maximizers is still pretty disappointing; and the AI risk discourse does not, traditionally, care about the "existing ecosystem" per se. But maybe it should care more? At the least, I think the "notkilleveryone" part of AI safety—that is, the part concerned with the AIs violating our boundaries, rather than with making sure that unclaimed galactic resources get used optimally—has resonance with "protect the existing ecosystem" vibes. And part of the problem with dictators, and with top-down-gone-wrong, is that some of the virtues of an ecosystem get lost.



Maybe we could do, like, ecosystem-onium? (Image source here.)

Yet for all that AI safety might seem to want more green out of the invention of AGI, I think it also struggles to coherently conceptualize what green even *is*. Indeed, I think that various strands of the AI safety literature can be seen as attempting to somehow formalize the sort of green we intuitively want out of our AIs. "Surely it's possible," the thought goes, "to build a powerful mind that doesn't want exactly what we want, but which also doesn't just drive the universe off into some extreme and valueless tail? Surely, it's possible to just, you know, not optimize that hard?" See, e.g., the literature on "soft optimization," "corrigibility," "low impact agents," and so on.[139] As far as I can tell, Yudkowsky has broadly declared defeat on this line of research,[140] on the grounds that vibes

[139] See e.g. this piece by Scott Garrabrant, characterizing such concepts as "green." Thanks to Daniel Kokotajlo for flagging.

[140] He has also declared defeat on all technical AI safety research, at least at current levels of human intelligence—"Nate and Eliezer both believe that humanity should not be attempting technical alignment at its current level of cognitive ability…" But the reason in this case is more specific.

of this kind are "anti-natural" to sufficiently smart agents that also get-things-done.[141] But this sounds a lot like saying: "sorry, the sort of green we want, here, just isn't enough of a coherent *thing*." And indeed: maybe not.[142] But if, instead, the problem is a kind of "green-blindness," rather than green-incoherence—a problem with the way a certain sort of philosophy blots out green, rather than with green itself—then the connection between green and AI safety suggests value in learning-to-see.

And I think green-blindness matters, too, because green is part of what protests at the kind of power-seeking that ideologies like rationalism and effective altruism can imply, and which warns of the dangers of *yang*-gone-wrong. Indeed, Yudkowsky's Lord of Dark, in dismissing green with contempt, also appears, notably, to be putting himself in a position to take over the world. There is no equilibrium, no balance of Nature, no God-to-be-trusted; instead there is poverty and pain and disease, too much to bear; and only nothingness above. And so, conclusion: cast the spell of Ultimate Power, young sorcerer. The universe, it seems, needs to be *controlled.*

And to be clear, in case anyone missed it: the Spell of Ultimate Power is a metaphor for AGI. The Lord of Dark is one of Yudkowsky's "programmers" (and one of Lewis's "conditioners"). Indeed, when the pain of the world breaks into the consciousness of the protagonist of the story, it does so in a manner extremely reminiscent of the way it breaks into young-Yudkowsky's consciousness, in his accelerationist days, right before he declares "*reaching the Singularity as fast as possible* to be the Interim Meaning of Life, the temporary definition of Good, and the foundation until further notice of my ethical system." (Emphasis in the original.)

> I have had it. I have had it with crack houses, dictatorships, torture chambers, disease, old age, spinal paralysis, and world hunger. I have had it with a death rate of 150,000 sentient beings per day. I have had it with this planet. I have had it with mortality. None of this is necessary. The time has come to stop turning away from the mugging on the corner, the beggar on the street. It is no longer necessary to close our eyes, blinking away the tears, and repeat the mantra: "I can't solve all the problems of the world." We *can*. We can *end* this.

Of course, young-Yudkowsky has since aged. Indeed, older-Yudkowsky has disavowed all of his pre-2002 writings, and he wrote that in 1996. But he wrote the Sword of the Good in 2009, and the protagonist, in that story, reaches a similar conclusion. At the request of the Lord of Dark, whose Spell of Ultimate Power requires the sacrifice of a wizard, the protagonist kills the wizard who warned about disrupting equilibrium, and gives his

[141] From "List of Lethalities": "**Corrigibility is anti-natural to consequentialist reasoning**; 'you can't bring the coffee if you're dead' for almost every kind of coffee. We (MIRI) tried and failed to find a coherent formula for an agent that would let itself be shut down (without that agent actively trying to get shut down). Furthermore, many anti-corrigible lines of reasoning like this may only first appear at high levels of intelligence... The second course is to build corrigible AGI which doesn't want exactly what we want, and yet somehow fails to kill us and take over the galaxies despite that being a convergent incentive there... The second thing looks unworkable (less so than CEV, but still lethally unworkable) because **corrigibility runs *actively counter* to instrumentally convergent behaviors** within a core of general intelligence (the capability that generalizes far out of its original distribution). You're not trying to make it have an opinion on something the core was previously neutral on. You're trying to take a system implicitly trained on lots of arithmetic problems until its machinery started to reflect the common coherent core of arithmetic, and get it to say that as a special case 222 + 222 = 555..."

[142] Though here and elsewhere, I think Yudkowsky overrates how much evidence "MIRI tried and failed to solve X problem" provides about X problem's difficulty.

sword—the Sword of the Good, which "kills the unworthy with a slightest touch" (but which only tests for intentions)—to the Lord of Dark to touch. "*Make it stop. Hurry*," says the protagonist. The Lord of Dark touches the blade and survives, thereby proving that his intentions are good. "I won't trust myself," he assures the protagonist. "I don't trust you either," the protagonist replies, "but I don't expect there's anyone better." And with that, the protagonist waits for the Spell of Ultimate Power to foom, and for the world as he knows it to end.

Is that what choosing Good looks like? Giving Ultimate Power to the well-intentioned—but un-accountable, un-democratic, Stalin-ready—because everyone else seems worse, in order to remake reality into something-without-darkness as fast as possible? And killing people on command in the process, without even asking why it's necessary, or checking for alternatives?[143] The story wants us, rightly, to approach the moral narratives we're being sold with skepticism; and we should apply the same skepticism to the story itself.

Perhaps, indeed, Yudkowsky aimed intentionally at prompting such skepticism (though the Lord of Dark's object-level schtick—his concern for animals, his interest in cryonics, his desire to tear-apart-the-foundations-of-reality-and-remake-it-new—seems notably in line with Yudkowsky's own). At the least, elsewhere in his fiction (e.g., HPMOR), he urges more caution in responding to the screaming pain of the world;[144] and his more official injunction towards "programmers" who have suitably solved alignment—i.e., "implement present-day humanity's coherent extrapolated volition"—involves, at least, certain kinds of inclusivity. Plus, obviously, his current, real-world policy platform is heavily *not* "build AGI as fast as possible." But as I've been emphasizing throughout this series, his underlying philosophy and metaphysics *is*, ultimately, heavy on the need for certain kinds of *control*; the need for the universe to be *steered*, and by the right hands; bent to the right will; mastered. And here, I think, green objects.

5 Green, according to non-Green

> *"Roofless, floorless, glassless, 'green to the very door'..."*
>
> *—Zadie Smith*

But what exactly is green's objection? And should it get any weight?

There's a familiar story, here, which I'll call "green-according-to-blue." On this story, green is worried that non-green is going to do *blue* wrong—that is, act out of inadequate *knowledge*. Non-green *thinks* it knows what it's doing, when it attempts to remake Nature in its own image (e.g. remaking the ecosystem to get rid of wild animal suffering)—but according to green-according-to-blue, it's overconfident; the system it's trying to steer is too complex and unpredictable. So thinks blue, in steel-manning green. And blue, similarly,

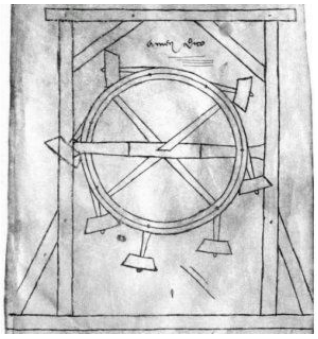
[143] Thanks to Arden Koehler for discussion, years ago.

[144] For example, his Harry Potter turns down the phoenix's invitation to destroy Azkaban, and declines to immediately give-all-the-muggles-magic, lest doing so destroy the world (though this latter move is a reference to the vulnerable world, and in practice, ends up continuing to concentrate power in Harry's hands).

talks about Chesterton's fence—about the status quo often having a reason-for-being-that-way, even if that reason is hard to see; and about approaching it with commensurate respect and curiosity. Indeed, one of blue's favored stories for mistrusting itself relies on deference to cultural evolution, and to organic, bottom-up forms of organization, in light of the difficulty of knowing-enough-to-do-better.

We can also talk about green according to something more like white. Here, the worry is that non-green will violate various moral rules in acting to reshape Nature. Not, necessarily, that it won't know what it's doing, but that what it's doing will involve trampling too much over the rights and interests of other agents/patients.

Finally, we can talk about green-according-to-black, on which green specifically urges us to accept things *that we're too weak to change*—and thus, to save on the stress and energy of trying-and-failing. Thus, black thinks that green is saying something like: don't waste your resources trying to build perpetual motion machines, or to prevent the heat death of the universe—you'll never be that-much-of-a-God. And various green-sounding injunctions against e.g. curing death ("it's a part of life") sound, to black, like mistaken applications (or: confused reifications) of this reasoning.[145]



Early design for a perpetual motion machine

I think that green does indeed care about all of these concerns—about ignorance, immorality, and not-being-a-God—and about avoiding the sort of straightforward mistakes that blue, white, and black would each admit as possibilities. Indeed, one way of interpreting green is to simply read it as a set of heuristics and reminders and ways-of-thinking that other colors do well, on their own terms, to keep in mind—e.g., a vibe that helps blue remember its ignorance, black its weakness, and so on. Or at least, one might think that this interpretation is what's *left over*, if you want to avoid attributing to green various crude naturalistic fallacies, like "everything Natural is Good," "all problems stem from human agency corrupting Nature-in-Harmony," and the like.[146]

But I think that even absent such crude fallacies, green-according-to-green has more to

[145]There's also a different variant of green-according-to-black, which urges us to notice the *power* of various products-of-Nature —for example, those resulting from evolutionary competition. Black is down with this—and down with competition more generally.

[146]Here I think of conversations I've had with utilitarian-ish folks, in which their attempts to fit environmentalism within their standard ways of thinking have seemed to me quite distorting of its vibe. "Is it kind of like: they think that ecosystems are moral patients?" "Is it like: they want to maximize Nature?"

add to the other colors than this. And I think that it's important to try to really grok what it's adding. In particular: a key aspect of Yudkowsky's vision, at least, is that the ignorance and weakness situation is going to alter dramatically post-AGI. Blue and black will foom hard, until earth's future is chock full of power and knowledge (even if also: paperclips). And as blue and black grow, does the need for green shrink? Maybe somewhat. But I don't think green itself expects obsolescence—and some parts of my model of green think that people with the power and science of transhumanists (and especially: of Yudkowskian "programmers," or Lewisian "conditioners") need the virtues of green all the more.

But what are those virtues? I won't attempt any sort of exhaustive catalog here. But I do want to try to point at a few things that I think the green-according-to-non-green stories just described might miss. Green cares about ignorance, immorality, and not-being-a-God—yes. But it also cares about them in a distinctive way—one that more paradigmatically blue, white, and black vibes don't capture very directly. In particular: I think that green cares about something like *attunement*, as opposed to just knowledge in general; about something like *respect*, as opposed to morality in general; and about taking a certain kind of joy in the dance of both *yin* and *yang*—in encountering an Other that is not fully "mastered"—as opposed to wishing, always, for fuller mastery.

I'll talk about attunement in my next essay—it's the bit of green I care about most. For now, I'll give some comments on respect, and on taking joy in both *yin* and *yang*.

6 Green and respect

In Being Nicer than Clippy, I tried to gesture at some hazy distinction between what I called "paperclippy" modes of ethical conduct, and some alternative that I associated with "liberalism/boundaries/niceness." Green, I think, tends to be fairly opposed to "paperclippy" vibes, so on this axis, a green ethic fits better with the liberalism/boundaries/niceness thing.

But I think that the sort of "respect" associated with green goes at least somewhat further than this—and its status in relation to more familiar notions of "Morality" is more ambiguous. Thus, consider the idea of casually cutting down a giant, ancient redwood tree for use as lumber—lifting the chainsaw, watching the metal bite into the living bark. Green, famously, protests at this sort of thing—and I feel the pull. When I stand in front of trees like this, they do, indeed, seem to have a kind of presence and dignity; they seem importantly *alive*.[147] And the idea of casual violation seems, indeed, repugnant.

[147] Albeit, one that it feels possible, also, to project onto many other life forms that we treat as much less sacred.

Albert Bierstadt's "Giant Redwood Trees of California" (Image source here).

But it remains, I think, notably unclear exactly how to fit the ethic at stake into the sorts of moral frameworks analytic ethicists are most comfortable with—including, the sort of rights-based deontology that analytic ethicists often use to talk about liberal and/or boundary-focused ethics.

Is the thought: the tree is instrumentally useful for human purposes? Environmentalists often reach for these justifications ("these ancient forests could hold the secret to the next vaccine"), but come now. Is that why people join the Sierra Club, or watch shows like Planet Earth? At the least, it's not what's on my own mind, in the forest, staring up at a redwood. Nor am I thinking "other people love/appreciate this tree, so we should protect it for the sake of their pleasure/preferences" (and this sort of justification would leave the question of *why* they love/appreciate it unelucidated).

Ok then, is the thought: the tree is beneficial to the welfare of a whole ecosystem of non-human moral-patient-y life forms? Again, a popular thought in environmentalist circles.[148] But again, not front-of-mind for me, at least, in encountering the tree itself; and in my mind, too implicating of gnarly questions about animal welfare and wild animal suffering to function as a simple argument for conservation.

Ok: is the thought, then, that the tree itself is a moral patient?[149] Well, kind of. The tree is *something*, such that you don't just do whatever you want with it. But again, in experiencing the tree as having "presence" or "dignity," or in calling it "alive," it doesn't feel like I'm also ascribing to it the sorts of properties we associate more paradigmatically with moral patience-y—e.g., consciousness. And talk of the tree as having "rights" feels strained.

And yet, for all this, something about just cutting down this ancient, living tree for lum-

[148]Though: the moral-patienthood question sometimes gets a bit fuzzed, for example re: ecosystems of plants.
[149]Or maybe, the ecosystem itself? See e.g. Aldo Leopold's "land ethic": "A thing is right when it tends to preserve the integrity, stability, and beauty of the biotic community. It is wrong when it tends otherwise."

ber does, indeed, feel pretty off to me. It feels, indeed, like some dimension related to "respect" is in deficit.

Can we say more about what this dimension consists in? I wish I had a clearer account. And it could be that this dimension, at least in my case, is just, ultimately, confused, or such that it would not survive reflection once fully separated from other considerations. Certainly, the arbitrariness of certain of the distinctions that some conservationist attitudes (including my own) tend to track (e.g., the size and age and charisma of a given life-form) raise questions on this front. And in general, despite my intuitive pull towards some kind of respect-like attitude towards the redwood, we're here nearby various of the parts of green that I feel most skeptical of.

It's because it's big isn't it... (Image source here.)

Still, before dismissing or reducing the type of respect at stake here, I think it's at least worth trying to bring it into clearer view. I'll give a few more examples to that end.

6.1 Blurring the self

I mentioned above that green is the "conservative" color. It cares about the past; about lineage, and tradition. If something life-like has survived, gnarled and battered and weathered by the Way of Things, then green often grants it more authority. It has had more harmonies with the Way of Things infused into it; and more disharmonies stripped away.

Of course, "harmony with the Way of Things" can be, just, another word for power (see also: "rationality"); and we can, indeed, talk about a lot of this in terms of blue and black—that is, in terms of the knowledge and strength that something's having-survived can indicate, even if you don't know what it is. But it can feel like the relationship green wants you to have with the past/lineage/tradition and so on goes beyond this, such that even if you actually get all of the power and knowledge you can out of the

past/lineage/tradition, you shouldn't just toss them aside. And this seems closely related to respect as well.

Part of this, I think, is that the past is a part of us. Or at least, our lineage is a part of us, almost definitionally. It's the pattern that created us; the harmony with the Way of Things that made us possible; and it continues to live within us and around us in ways we can't always see, and which are often well-worth discovering.

"Ok, but does that give it *authority* over us?" The quick straw-Yudkowskian answer is: "No. The thing that has authority over you, morally, is your heart; your *values*. The past has authority only insofar as some part of it is good according to those values."

But what if the past is *part of* your heart? Straw-Yudkowskianism often assumes that when we talk about "your values," we are talking about something that lives *inside you*; and in particular, mostly, inside your brain. But we should be careful not to confuse the brain-as-seer and the brain-as-thing-seen. It's true that ultimately, your brain moves your muscles, so anything with the sort of connection to your behavior adequate to count as "your values" needs to get some purchase on your brain somehow. But this doesn't mean that your brain, in seeking out guidance about what to do, needs to look, ultimately, *to itself*. Rather, it can look, instead, outwards, towards the world. "Your values" can make essential reference to bits of Reality beyond yourself, that you cannot see directly, and must instead discover—and stuff about your past, your lineage, and so on is often treated as a salient candidate for mattering in this respect; an important part of "who you are."

MOANA song "We Know The Way" (link to video). See also this one.

In this way, your "True Self" can be mixed-up, already, with that strange and unknown Other, reality. And when you meet that Other, you find it, partly, as mirror. But: the sort of mirror that shows you something you hadn't seen before. Mirror, but also window.

Green, traditionally, is interested in these sorts of line-blurrings—in the ways in which it might not be me-over-here, you-over-there; the way the real-selves, the true-Agents, might be more spread out, and intermixed. Shot through forever with each other. Until, in the limit, it was God the whole time: waking up, discovering himself, meeting himself in each other's eyes.

Of course, God does, still, sometimes need to go to war with parts himself—for example, when those parts are invading Poland. Or at least, *we* do—for our true selves are not, it seems, God entire; that's the "evil" problem. But such wars need not involve saying "I see none of myself in you." And indeed, green is very wary of stances towards evil and

darkness that put it, too much, "over there," instead of finding ourselves in its gaze. This is a classic Morality thing, a classic failure mode of White. But green-like lessons often go the opposite direction. See, for example, the Wizard of Earthsea, or the ending of Moana (spoilers at link). Your true name, perhaps, lies partly in the realm of shadow. You can still look on evil with defiance and strength; but to see fully, you must learn to look in some other way as well.

And here, perhaps, is one rationale for certain kinds of respect. It's not, just, that something that might carry knowledge and power you can acquire and use, or fear; or that it might conform to and serve some pre-existing value you know, already, from inside yourself. Rather, it might also carry some part of your heart itself inside of it; and to kill it, or to "use it," or put it too much "over there," might be to sever your connection with your whole self; to cut some vein, and so become more bloodless; to block some stream, and so become more dry.

6.2 Respecting superintelligences

Moro the wolf God

I'll also mention another example of green-like "respect"—one that has more relevance to AI risk.

Someone I know once complained to me that the Yudkowsky-adjacent AI risk discourse gives too little "respect" to superintelligences. Not just superintelligent AIs; but also, other advanced civilizations that might exist throughout the multiverse. I thought it was an interesting comment. Is it true?

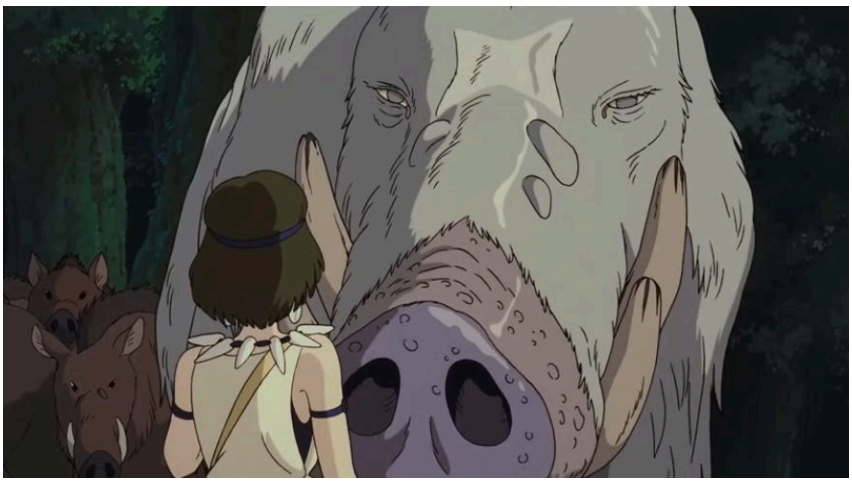
Certainly, straw-Yudkowskian-ism knows how to *positively appraise* certain traits possessed by superintelligences—for example, their smarts, cunning, technological prowess, etc (even if not also: their values). Indeed, for whatever notion of "respect" one directs at a formidable adversary trying to kill you, Yudkowsky seems to have a lot of *that* sort of respect for misaligned AIs. And he worries that our species has too little.

That is: Yudkowsky respects the *power* of superintelligent agents. And he's generally

happy, as well, to respect their moral rights. True, as I discussed in "Being nicer than Clippy," I do think that the Yudkowskian AI risk discourse sometimes under-emphasizes various key aspects of this. But that's not what I want to focus on here.

Once you've positively appraised the power (intelligence, oomph, etc) of a superintelligent agent, though, and given its moral claims adequate weight, what bits are left to respect? On a sufficiently abstracted Yudkowskian ontology, the most salient candidate is just: the utility function bit (agents are just: utility functions + power/intelligence/oomph). And sure, we can positively appraise utility functions (and: parts of utility functions), too—especially to the degree that they are, you know, *like ours.*

But some dimension of respect feels like it might be missing from this picture. For one thing: real world creatures—including, plausibly, quite oomph-y ones—aren't, actually, combinations of utility functions and degrees-of-oomph. Rather, they are something more gnarled and detailed, with their own histories and cultures and idiosyncrasies—the way the boar god smells you with his snout; the way humans are quiet at funerals; the way ChatGPT was trained to predict the human internet. And respect needs to attend to and adjust itself to a creature's contours—to craft a specific sort of response to a specific sort of being. Of course, it's hard to do that without *meeting* the creature in question. But when we view superintelligent agents centrally through the lens of rational-agent models, it's easy to forget that we should do it at all.



Okkoto the blind boar God

But even beyond this need for specificity, I think some other aspect of respect might be missing too. Suppose, for example, that I meet a super-intelligent emissary from an ancient alien civilization. Suppose that this emissary is many billions of years old. It has traveled throughout the universe; it has fought in giant interstellar wars; it understands reality with a level of sophistication I can't imagine. How should I relate to such a being?

Obviously, indeed, I should be scared. I should wonder about what it can do, and what it wants. And I should wonder, too, about its moral claims on me. But beyond that, it seems appropriate, to me, to approach this emissary with some more holistic humility and open

attention. Here is an ancient demi-God, sent from the fathoms of space and time, its mind tuned and undergirded by untold depths of structure and harmony, knowledge and clarity. In a sense, it stands closer to reality than we do; it is a more refined and energized expression of reality's nature, pattern, Way. When it speaks, more of reality's voice speaks through it. And reality sees more truly through its eyes.

Does that make it "good"? No—that's the orthogonality thing, the AI risk thing. But it likely has much more of whatever "wisdom" is compatible with the right ultimate picture of "orthogonality"—and this might, actually, be a lot. At the least, insofar as we are specifically trying to get the "respect" bit (as opposed to the not-everyone-dying bit) right, I worry a bit about coming in too hard, at the outset, with the conceptual apparatus of orthogonality; about trying, too quickly, to carve up this vast and primordial Other Mind into "capabilities" and "values," and then taking these carved-up pieces, centrally, as objects of positive or negative appraisal.

In particular: such a stance seems notably loaded on our standing in judgment of the super-intelligent Other, according to our own pre-existing concepts and standards; and notably lacking on interest in the Other's judgment of us; or in understanding the Other on its own terms, and potentially growing/learning/changing in the process. Of course, we should still do the judging-according-to-our-own-standards bit—not to mention, the not-dying bit. But shouldn't we be doing something else as well?

Or to put it another way: faced with an ancient super-intelligent civilization, there is a sense in which we humans are, indeed, as children.[150] And there is a temptation to say we should be acting with the sort of holistic humility appropriate to children vis-à-vis adults—a virtue commonly associated with "respect."[151] Of course, some adults are abusive, or evil, or exploitative. And the orthogonality thing means you can't just trust or defer to their values either. Nor, even in the face of superintelligence, should we cower in shame, or in worship—we should stand straight, and look back with eyes open. So really, we need the virtues of children who are respectful, *and* smart, *and* who have their own backbone—the sort of children who manage, despite their ignorance and weakness, to navigate a world of flawed and potentially threatening adults; who become, quickly, adults themselves; and who can hold their own ground, when it counts, in the meantime. Yes, a lot of the respect at stake is about the fact that the adults are, indeed, smarter and more powerful, and so should be feared/learned-from accordingly. But at least if the adults meet *certain* moral criteria—restrictive enough to rule out the abusers and exploiters, but not so restrictive as to require identical values—then it seems like green might well judge them worthy of some other sort of "regard" as well.

But even while it takes some sort of morality into account, the regard in question also seems importantly distinct from direct moral approval. Here I think again of Miyazaki movies, which often feature creatures that mix beauty and ugliness, gentleness and violence; who seem to live in some moral plane that intersects and interacts with our own, but

[150] Thanks to Nick Bostrom for discussion of this a while ago.

[151] See also Bostrom re: our interactions with superintelligent civilizations: 'We should be modest, willing to listen and learn. We should not too headstrongly insist on having too much our way. Instead, we should be compliant, peace-loving, industrious, and humble…" Though I have various questions about his picture in that paper.

which moves our gaze, too, along some other dimension, to some unseen strangeness.[152] Wolf gods; blind boar gods; spirits without faces; wizards building worlds out of blocks marred by malice—how do you live among such creatures, and in a world of such tragedy and loss? "I am making this movie because I do not have the answer," says the director, as he bids his art goodbye.[153] But some sort of respect seems apt in many cases—and of a kind that can seem to go beyond "you have power," "you are a moral patient," and "your values are like mine."

I admit, though, that I haven't been able to really pin down or elucidate the type of respect at stake.[154] In the appendix to this essay, I discuss one other angle on understanding this sort of respect, via what I call "seeking guidance from God." But I don't feel like I've nailed that angle, either—and the resulting picture of green brings it nearer to "naturalistic fallacies" I'm quite hesitant about. And even the sort of respect I've gestured at in the examples above—for trees, lineages, superintelligent emissaries, and so on—risks various types of inconsistency, complacency, status-quo-bias, and getting-eaten-by-aliens. And perhaps it cannot, ultimately, be made simultaneously coherent and compelling.

But I feel some pull in this direction all the same. And regardless of our ultimate views on this sort of respect, I think it's not quite the same thing as e.g. making sure you respect Nature's "rights," or conform to the right "rules" in relation to it—what I called, above, "green-according-to-white."

7 Green and joy

> *"Pantheism is a creed not so much false as hopelessly behind the times. Once, before creation, it would have been true to say that everything was God. But God created: He caused things to be other than Himself that, being distinct, they might learn to love Him, and achieve union instead of mere sameness. Thus He also cast His bread upon the waters."*
>
> *—C.S. Lewis, in the Problem of Pain*

[152] Thanks to my sister, Caroline Carlsmith, for discussion.

[153] See quote here. Though: he's retired before as well...

[154] And this especially once we try to isolate out both the more directly morality-flavored bits, and the more power/knowledge-flavored bits—the sense in which green-like respect is caught up with trying to live, always, in a world, and amidst other agents and optimization processes, that you do not fully understand and cannot fully control. And indeed, perhaps part of what's going on here is that green often resists attempts to re-imagine our condition without—or even, with substantially less—of these constraints; to ask questions like "Ok, but how would this attitude alter if you instead had arbitrary knowledge and power?" Green, one suspects, is skeptical of hypotheticals like this; they seem, to green, like too extreme a departure from who-we-are, where-we-live. Part of this may be that familiar "I refuse to do thought experiments that would isolate different conceptual variables" thing that so frustrates philosophers, and which so stymies attempts to clarify and pull apart different concepts. But I wonder if there is some other wisdom—related, perhaps, to just how deeply our minds are *for* not-knowing, not-having-full-control—in play.

"The ancient of days" by William Blake (Image source here; strictly speaking this isn't God but whatever...)

I want to turn, now, to green-according-to-black, according to which green is centrally about recognizing our ongoing weakness—just how much of the world is not (or: not yet) master-able, controllable, *yang*-able.

I do think that *something* in the vicinity is a part of what's going on with green. And not just in the sense of "accepting things you can't change." Even if you *can* change them, green is often hesitant about attempting forms of change that involve lots of effort and strain and *yang*. This isn't to say that green doesn't do anything. But when it does, it often tries to find and ride some pre-existing "flow"—to turn keys that fit easily into Nature's locks; to guide the world in directions that it is fairly happy to go, rather than forcing it into some shape that it fights and resists.[155] Of course, we can debate the merits of green's priors, here, about what sorts of effort/strain are what sorts of worth it—and indeed, as mentioned, green's tendency towards unambition and passivity is one of my big problems with it. But everyone, even black, agrees on the merits of energy efficiency; and in the limit, if *yang* will definitely fail, then *yin* is, indeed, the only option. Sad, says black, but sometimes necessary.

Here, though, I'm interested in a different aspect of green—one which does not, like black,

[155]Thanks to Anna Salamon for some discussion here.

mourn the role of *yin*; but rather, takes joy in it. Let me say more about what I mean.

7.1 Love and otherness

"I have bedimm'd
The noontide sun, call'd forth the mutinous winds,
And 'twixt the green sea and the azured vault
Set roaring war: to the dread rattling thunder
Have I given fire and rifted Jove's stout oak
With his own bolt; the strong-based promontory
Have I made shake and by the spurs pluck'd up
The pine and cedar: graves at my command
Have waked their sleepers, oped, and let 'em forth
By my so potent art..."

—Prospero

"Scene from Shakespeare's The Tempest," by Hogarth (Image source here)

There's an old story about God. It goes like this. First, there was God. He was pure *yang*, without any competition. His was the Way, and the Truth, and the Light—and no else's. But, there was a problem. He was too alone. Some kind of "love" thing was too missing.

So, he created Others. And in particular: *free* Others. Others who could turn to him in love; but also, who could turn away from him in sin—who could be what one might call "misaligned."

And oh, they were misaligned. They rebelled. First the angels, then the humans. They became snakes, demons, sinners; they ate apples and babies; they hurled asteroids and lit the forests aflame. Thus, the story goes, evil entered a perfect world. But somehow, they say, it was in service of a higher perfection. Somehow, it was all caught up with the possibility of love.

The Fall of the Rebel Angels, by Bosch. (Image source here.)

Why do I tell this story? Well: a lot of the "deep atheism" stuff, in this series, has been about the problem of evil. Not, quite, the traditional theistic version—the how-can-God-be-good problem. But rather, a more generalized version—the problem of how to relate, spiritually, to an orthogonal and sometimes horrifying reality; how to live in the light of one's vulnerability to an unaligned God. And I've been interested, in particular, in responses to this problem that focus, centrally, on reducing the vulnerability in question—on seeking greater power and control; on "deep atheism, therefore black." These responses attempt to reduce the share of the world that is Other, and to make it, instead, a function of Self (or at least, the self's heart). And in the limit, it can seem like they aspire (if only it were possible) to abolish the Other entirely; to control *everything*, lest any evil or misalignment sneak through; and in this respect, to take up that most ancient and solitary throne—one that God sat on, before the beginning of time; the throne of pure *yang*.

So I find it interesting that God, in the story above, rejected this throne. Unlike us, he had the option of full control, and a perfectly aligned world. But he chose something different. He left pure self behind, and chose instead to create Otherness—and with it, the possibility (and reality) of evil, sin, rebellion, and all the rest.

Of course, we might think he chose wrong. Indeed, the story above is often offered as a defense (the "free will defense") of God's goodness in the face of the world's horrors—and we might, with such horrors vividly before us, find such a defense repugnant.[156] At the least, couldn't God have found a better version of freedom? And one might worry, too, about the metaphysics of the freedom implicitly at stake. In particular, at least as Lewis tells it,[157] the story loads, centrally, on the idea that instead of determining the values of his creatures (and without, one assumes, simply *randomizing* the values that they get, or letting some other causal process decide), God can just give them freedom instead—the freedom to have some part of them uncreated; to be an uncaused cause. But in our naturalistic universe, at least, and modulo various creative theologies, this doesn't seem

[156] And this even setting aside the other philosophical problems with such a move.
[157] Let's set Calvin aside.

like something a creator (especially an omniscient and omnipotent one) can do. Whether his creatures are aligned, or unaligned, God either made them so, or he let some other not-them process (e.g., his random-number-generator) do the making. And once we've got a better and more compatibilist metaphysics in view, the question of "why not make them *both* good *and* free?" becomes much more salient (see e.g. my discussion of Bob the lover-of-joy here). And note, importantly, that the same applies to us, with our AIs.[158]

But regardless of how we feel about God's choice in the story, or the metaphysics it presumes, I think it points at something real: namely, that we don't, actually, always want more power, control, *yang*. To the contrary, and even setting aside more directly ethical constraints on seeking power over others, a lot of our deepest values are animated by taking certain kinds of *joy* in otherness and *yin*—in being not-God, and relatedly: not-alone.

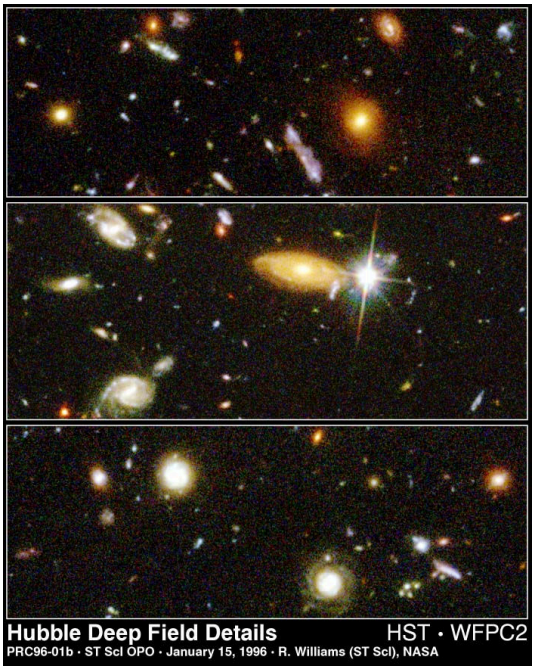
Love is indeed the obvious example here. Love, famously, is directed (paradigmatically) at something outside yourself—something present, but exceeding your grasp; something that surprises you, and dances with you, and look back at you. True, people often extoll the "sameness" virtues of love—unity, communion, closeness. But to merge, fully—to make love centrally a relation with an (expanded) self—seems to me to miss a key dimension of joy-in-the-Other per se.

Here I think of Martin Buber's opposition, in more spiritual contexts, to what he calls "doctrines of immersion" (Buddhism, on his reading, is an example), which aspire to *dissolve* into the world, rather than to *encounter* it. Such doctrines, says Buber, are "based on the gigantic delusion of human spirit bent back into itself—the delusion that spirit occurs in man. In truth it occurs from man—between man and what he is not."[159] Buber's spirituality focuses, much more centrally, on this kind of "between"—and compared with spiritual vibes focused more on unification, I've always found his vision the more resonant. Not to merge, but stand face to face. Not *become* the Other; but to speak, and to listen, in dialogue. And many other interpersonal pleasures—conversation, friendship, community—feature this kind of "between" as well.

Or consider experiences of wonder, sublimity, beauty, curiosity. These are all, paradigmatically, experiences of encountering or receiving something outside yourself—something that draws you in, stuns you, provokes you, overwhelms you. They are, in this sense, a type of *yin*. They *discover* something, and take joy in the discovery. Reality, in such experiences, is presented as electric and wild and alive.

[158] That is, there is no alternative to alignment like "just let the AIs be an uncaused-cause of their own values." Either we will create their values, or some other process will.

[159] See quote here.



(Image source here)

And many of the activities we treasure specifically involve a play of *yin* and *yang* in relation to some not-fully-controlled other—consider partner dancing, or surfing, or certain kinds of sex. And of course, sometimes we go to an activity seeking the *yin* bit in particular. Cf, e.g., dancing with a good lead, sexual submissiveness, or letting a piece of music carry you.

"Dance in the country," by Renoir. (Image source here.)

And no wonder that our values are like this. Humans are extremely not-Gods. We evolved

in a context in which we had, always, to be learning from and responding to a reality very much beyond-ourselves. It makes sense, then, that we learned, in various ways, to take joy in this sort of dance—at least, sometimes.

Still, especially in the context of abstract models of rationality that can seem to suggest a close link between being-an-agent-at-all and a voracious desire for power and control, I think it's important to notice how thoroughly joy in various forms of Otherness pervades our values.[160] And think this joy is at least one core thing going on with green. Contra green-according-to-black, green isn't just *resigned* to *yin*, or "serene" in the face of the Other. Green *loves* the Other, and gets excited about God. Or at least, God in certain guises. God like a friend, or a newborn bird, or a strange and elegant mathematical pattern, or the cold silence of a mountain range. God qua object of wonder, curiosity, reverence, gentleness. True, not all God-guises prompt such reactions—cancer, the Nazis, etc are still, more centrally, to-be-defeated.[161] But contra Black (and even modulo White), neither is everything either a matter of mastery, or of too-weak-to-win.

(Image source here.)

7.2 The future of *yin*

"But this rough magic
I here abjure, and, when I have required
Some heavenly music, which even now I do,
To work mine end upon their senses that
This airy charm is for, I'll break my staff,
Bury it certain fathoms in the earth,
And deeper than did ever plummet sound
I'll drown my book."

—Prospero

[160] Indeed, in many cases, I think it's not even clear what total power and control would even mean—see e.g. Grace's "total horse takeover" for some interestingly nuanced analysis.

[161] Though to-be-defeated is compatible with to-be-loved.

What's more, I think this aspect of our values actually comes under threat, in the age of AGI, from a direction quite different from the standard worry about AI risk. The AI risk worry is that we'll end up with too little *yang* of our own, at least relative to some Other. But there is another, different worry—namely, that we'll end up with too *much* yang, and so lose various of the joys of Otherness.

It's a classic sort of concern about Utopia. What does life become, if most aspects of it can be chosen and controlled? What is love if you can design your lover? Where will we seek wildness if the world has been tamed? Yudkowsky has various essays on this; and Bostrom has a full book shortly on the way. I'm not going to try to tackle the topic in any depth here—and I'm generally skeptical of people who try to argue, from this, to Utopia not being *extremely better,* overall, than our present condition. But just because Utopia is better *overall* doesn't mean that nothing is lost in becoming able to create it—and some of the joys of *yin* (and relatedly, of *yang*—the two go hand in hand) do seem to me to be at risk. Hopefully, we can find a way to preserve them, or even deepen them.[162] And hopefully, while still using the future's rough magic wisely, rather than breaking staff and drowning book.

Still, I wonder where a wise and good future might, with Prospero, abjure certain alluring sorceries—and not just for lack of knowledge of how they might shake the world. Where the future might, with Ogion, let the rain fall. At the least, I find interesting the way various transhumanist visions of the future—what Ricón (2021) calls "cool sci-fi-shit futurism"—often read as cold and off-putting precisely insofar as they seem to have lost touch with some kind of green. Vibes-wise—but also sometimes literally, in terms of color-scheme: everything is blue light and chrome and made-of-computers. But give the future green—give it *plants*, fresh air, mountain-sides, sunlight—and people begin to warm to Utopia. Cf. solarpunk, "cozy futurism," and the like. And no wonder: green, I think, is closely tied with many of our most resonant visions of happiness.

Example of solarpunk aesthetic (to be clear: I think the best futures are *way* more future-y than this)

[162] And in some cases, I think the sense of threat comes from a clearer vision of the universe as mechanistic and predictable, rather than from something having more fundamentally *changed*.

Maybe, on reflection, we'll find that various more radical changes are sufficiently better that it's worth letting go of various more green-like impulses—and if so, we shouldn't let conservatism hold us back. Indeed, my own best guess is that a lot of the value lies, ultimately, in this direction, and that the wrong sort of green could lead us catastrophically astray. But I think these more green-like visions of the future actually provide a good starting point, in connecting with the possible upsides of the Utopia. Whichever direction a good future ultimately grows, its roots will have been in our present loves and joys—and many of these are green.

Alexander speaks about the future as a garden. And if a future of nano-bot-onium is pure *yang*, pure top-down control, gardens seem an interesting alternative—a mix of *yin* and *yang*; of your work, and God's, intertwined and harmonized. You seed, and weed, and fertilize. But you also let-grow; you let the world *respond*. And you take joy in what blooms from the dirt.

8 Next up: Attunement

Ok, those were some comments on "green-according-to-white," which focuses on obeying the right moral rules in relation to Nature, and "green-according-black," which focuses on accepting stuff that you're too weak to change. In each case, I think, the relevant diagnosis doesn't quite capture the full green-like thing in the vicinity, and I've tried to do at least somewhat better.

But I haven't yet discussed "green-according-to-blue," which focuses on making sure we don't act out of inadequate knowledge. This is probably the most immediately resonant reconstruction of green, for me—and the one closest to the bit of green I care most about. But again, I think that blue-like "knowledge," at least in its most standard connotation, doesn't quite capture the core thing—something I'll call "attunement." In my next essay, I'll say more about what I mean.

9 Appendix: Taking guidance from God

This appendix discusses one other way of understanding the sort of "conservatism" and "respect" characteristic of green—namely, via the concept of "taking guidance from God." This is a bit of green that I'm especially hesitant about, and I don't think my discussion nails it down very well. But I thought I would include some reflections regardless, in case they end up useful/interesting on their own terms.

Earlier in the series, I suggested that "deep atheism" can be seen, fundamentally, as emerging from severing the connection between *Is* and *Ought*, the *Real* and the *Good*. Traditional theism can trust that somehow, the two are intimately linked. But for deep atheism, they become orthogonal—at least conceptually.[163] Maaayybe some particular *Is*

[163] And it's this same orthogonality that kills you, when the *Is* gets amped-up-to-foom via bare intelligence.

is *Ought*; but only contingently so—and on priors, probably not.[164] Hence, indeed, deep atheism's sensitivity to the so-called "Naturalistic fallacy," which tries to move illicitly from *Is* to *Ought*, from *Might* in the sense of "strong enough to exist/persist/get-selected" to *Right* in the sense of "good enough to seek guidance from." And naturalistic fallacies are core to deep atheism's suspicion towards green. Green, the worry goes, seeks too much input from God.

What's more, I think we can see an aspiration to "not seek input from God" in various other more specific ethical motifs associated with deep atheist-y ideologies like effective altruism. Consider, for example, the distinction between doing and allowing, or between action and omission.[165] Consequentialism—the ethical approach most directly associated with Effective Altruism—is famously insensitive to distinctions like this, at least in theory. And why so? Well, one intuitive argument is that such distinctions require treating the "default path"—the world that results if you go fully *yin*, if you merely *allow* or *omit*, if you "let go and let God"—as importantly different from a path created by your own *yang*. And because God (understood as the beyond-your-*yang*) sets the "default," ascribing intrinsic importance to the "default" is already to treat God's choice as ethically interesting—which, on deep atheism, it isn't.[166]

Worse, though: distinctions like acts vs. omissions and doing vs. allowing generally function to *actively defer* to God's choice, by treating deviation from the "default" as subject to a notably *higher* burden of proof. For example, on such distinctions, it is generally thought much easier to justify letting someone die (for example, by not donating money; or in-order-to-save-five-more-people) than it is to justify killing them. But this sort of burden of proof effectively grants God a greater license-to-kill than it grants to the Self.[167] Whence such deference to God's hit list?

Or consider another case of not-letting-God-give-input: namely, the sense in which total utilitarianism treats *possible people* and *actual people* as ethically-on-a-par. Thus, in suitably clean cases, total utilitarianism will choose to create a new happy person, who will live for 50 years, rather than to extend an existing happy human's life by another 40 years. And in combination with total-utilitarianism's disregard for distinctions like acts vs. omissions, this pattern of valuation can quickly end up *killing* existing people in order to replace them with happier alternatives (this is part of what gives rise to the paperclipping problems I discussed in "Being nicer than Clippy"). Here, again, we see a kind of disregard-for-God's-input at work. An already-existing person is a kind of *Is*—a piece of the Real; a work of God.[168] But who cares about God's works? Why not bulldoze them and build something more optimal instead? Perhaps actual people have more *power* than possible people, due to already *existing*, which tends to be helpful from a power perspective. But

[164] At least if you're working with a conception of Goodness on which to be Good is to be what I previously called "a particular way."

[165] More on my take on this distinction here.

[166] You can pump this intuition even harder if you imagine that the default path in question was set via some source of randomness— e.g., a coin flip. H/t Cian Dorr for invoking this intuition in conversation years ago.

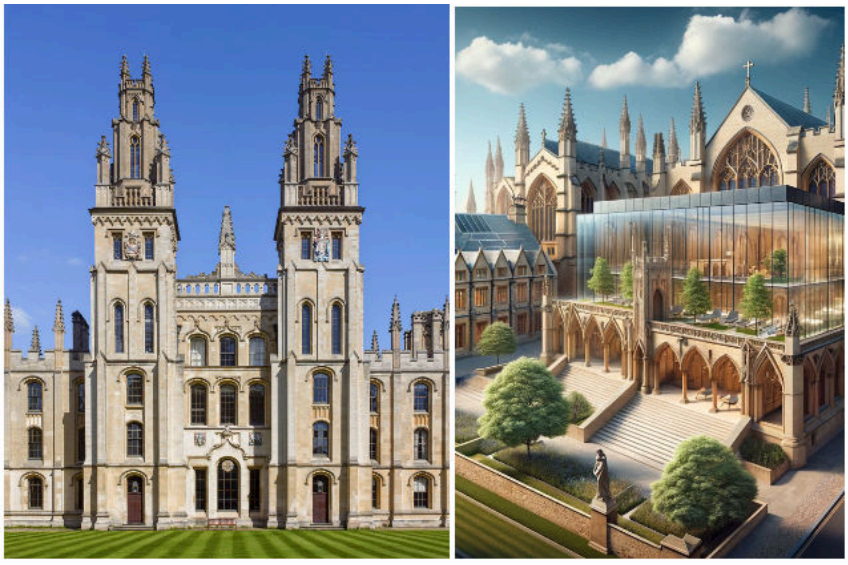
[167] Note that this includes God acting through the actions of others. That is, doing vs. allowing distinctions generally think that you can't e.g. kill one to prevent five others from being killed-by-someone-else; but that it is permissible to *let* one be killed-by-someone-else in order to prevent five people from being killed-by-someone-else.

[168] In principle you could've made the person with your own yang. But often not so.

a core *ethical* shtick, here, is about avoiding might-makes-right; about not taking moral cues from power alone. And absent might-makes-right, why does the fact that some actual-person *happens* to exist makes their welfare more important than that of those other, less-privileged possibilia?

Many "boundaries," in ethics, raise questions of this form. A boundary, typically, involves some work-of-God, some *Is* resulting from something other than your own *yang*. Maybe it's a fence around a backyard; or a border around a country; or a skin-bag surrounding some cells—and typically, *you* didn't build the fence, or found the country, or create the creature in question. God did that; Power did that. But from an *ethical* as opposed to a practical perspective, why should Power have a say in the matter? Thus, indeed, the paperclipper's atheism. Sure, OK: God loves the humans enough to have made-them-out-of-atoms (at least, for now). But Clippy does not defer to God's love, and wants those atoms for "something else." And as I discussed earlier in the series: utilitarianism reasons the same.

Or as a final example of an opportunity to seek or not-seek God's input, consider various flavors of what G.A. Cohen calls "small-c conservatism." According to Cohen, small-c conservatism is, roughly, an ethical attitude that wants to conserve existing valuable things—institutions, practices, ways of being, pieces of art—to a degree that goes above and beyond just wanting valuable things to exist. Here Cohen gives the example of All Souls College at Oxford University, where Cohen was a professor. Given the opportunity to tear down All Souls and replace it with something better, Cohen thinks we have at least some (defeasible) reason to decline, stemming just from the fact that All Souls already exists (and is valuable).[169] In this respect, small-c conservatism is a kind of ethical status quo bias—being already-chosen-by-God gives something an ethical leg up.[170]



Real All Souls on the left, ChatGPT-generated new version on the right. Though in the actual thought experiment ChatGPT's would be actually-better.

[169] Cohen *doesn't* think we have reason to preserve existing things that are *bad*.
[170] See Nebel (2015) for a defense of the rationality of status quo bias of this kind.

Various forms of environmental conservation, a la the redwoods above, are reminiscent of small-c conservative in this sense.[171] Consider, e.g., the Northern White Rhino. Only two left—both female, guarded closely by human caretakers, and unable to bear children themselves.[172] Why guard them? Sam Anderson writes about the day the last male, Sudan, died:

> We expect extinction to unfold offstage, in the mists of prehistory, not right in front of our faces, on a specific calendar day. And yet here it was: March 19, 2018. The men scratched Sudan's rough skin, said goodbye, made promises, apologized for the sins of humanity. Finally, the veterinarians euthanized him. For a short time, he breathed heavily. And then he died.
>
> The men cried. But there was also work to be done. Scientists extracted what little sperm Sudan had left, packed it in a cooler and rushed it off to a lab. Right there in his pen, a team removed Sudan's skin in big sheets. The caretakers boiled his bones in a vat. They were preparing a gift for the distant future: Someday, Sudan would be reassembled in a museum, like a dodo or a great auk or a Tyrannosaurus rex, and children would learn that once there had been a thing called a northern white rhinoceros.

Sudan's grave (Image source here)

Sudan's death went temporarily viral. And the remaining females are still their own attraction. People visit the enclosure. People cry for the species poached-to-extinction. Why the tears? Not, I think, from maybe-losing-a-vaccine. "At a certain point," writes Anderson, "we have to talk about love."

But what sort of love? Not the way the utilitarian loves the utilons. Not a love that mourns, equally, all the possible species that never got to exist—the fact that God created

[171] Though not always with a better alternative in the offing.

[172] My understanding is that the main options for saving the species involve (a) implanting fertilized eggs in another rhino sub-species or (b) something more Jurassic-park-y.

the Northern White Rhino in particular matters, here. No, the love at stake is more like: the way you love your dog, or your daughter, or your partner in particular. The way we love our languages and our traditions and our homes. A love that does more than compare-across-possibilia. A love that takes the *actual*, the *already*, as an input.

Of course, these examples of "taking God's guidance" are all different and complicated in their own ways. But to my mind, they point at some hazy axis along which one can try, harder and harder, to isolate the *Ought* from the influence of the *Is*. And this effort culminates in an attempt to stand, fully, outside of the world—the past, the status quo—so as to pass judgment on them all from some other, ethereal footing.

As ever, total utilitarianism—indeed, total-anything-ism—is an extreme example here. But we see the aesthetic of total utilitarianism's stance conjured by the oh-so-satisfying discipline of "population axiology" more generally—a discipline that attempts to create a function, a heart, that takes in all possible worlds (the actual world generally goes unlabeled), and spits out a consistent, transitive ranking of their goodness.[173] And Yudkowskians often think of their own hearts, and the hearts of the other player characters (e.g., the AIs-that-matter), on a similar model. Theirs isn't, necessarily, a ranking of *impartial* goodness; rather, it's a ranking of how-much-I-prefer-it, utility-according-to-me. But it applies to similar objects (e.g., possible "universe-histories"); it's supposed to have similar structural properties (e.g., transitivity, completeness, etc); and it is generated, most naturally, from a similar stance-beyond-the-world—a stance that treats you as a *judge* and a *creator* of worlds; and not, centrally, as a resident.[174] Indeed, from this stance, you can see all; you can compare, and choose, between anything.[175] All-knowing, all-powerful—it's a stance associated, most centrally, with God himself. Your heart, that is, is the "if I was God" part. No wonder, then, if it doesn't seek the real God's advice.[176]

[173] And ideally, a cardinal ranking that can then guide your choices between lotteries over such worlds.

[174] Even if your utility function makes essential reference to yourself, treating it as ranking "universe histories" requires looking at yourself from the outside.

[175] See here for an example of me appealing to this stance in the context of the von-Neumann Morgenstern utility theorem—one of the most common arguments for values needing to behave like utility functions: "Here's how I tend to imagine the vNM set-up. Suppose that you're hanging out in heaven with God, who is deciding what sort of world to create. And suppose, per impossible, that you and God aren't, in any sense,"part of the world." God's creation of the world isn't *adding* something to a pre-world history that included you and God hanging out; rather, the world is everything, you and God are deciding what kind of "everything" there will be, and once you decide, neither of you will ever have existed."

[176] Of course, it is possible to try to create "utility functions" that are sensitive to various types of input-from-the-real-God—to acts vs. omissions; to actual vs. possible people; to various existing boundaries and status-quos and endangered species and so on. Indeed, the Yudkowskians often speak about how rich and complicated their values are, while also, simultaneously, assuming that those values shake out, on reflection, into a coherent, transitive, cardinally-valued utility function (Since otherwise, their reflective selves would be executing a "dominated strategy," which it must be free to not do, right?). But if you hope to capture some distinction like acts vs. omissions or actual vs. possible people in a standard-issue utility function, while preserving at-least-decently your other intuitions about what matters and why, then I encourage you: give it an actual try, and see how it goes.

The philosophers, at least, tend to hit problems fast. The possible vs. actual people thing, for example, leads very quickly (in combination with a few other strong intuitions) to violations of transitivity and related principles (see e.g. the "Mere Addition" argument I discuss here; and Beckstead (2013), chapter 4); and the sort of deontological ethics most associated with acts vs. omissions, boundaries, and so on is rife with intransitivities and other not-very-utility-function-ish behavior as well (see e.g. this paper for some examples. Or try reading Frances Kamm, then see how excited you are about turning her views into a utility function over universe histories.) This isn't to say that you can't, ultimately, shoe-horn various forms of input-from-God into a consistent, ethically-intuitive utility function over all possible universe-histories (and some cases, I think, will be harder than others—See the literature on "consequentializing moral theories" for more on this—though not all "consequentializers" impose coherence constraints on the results of their efforts). But people rarely actually do the

But green-like respect, I think, often *does* seek God's advice. And more generally, I think, green's ethical motion feels less like ranking all possible worlds from ethereal stance-beyond, and then getting inserted into the world to move it up-the-ranking; and more like: lifting its head, looking around, and trying to understand and respond to what it sees.[177] After all: how did you learn, actually, what sorts of worlds you wanted? Centrally: by looking around the place where you are.

That said, not all of the examples of "taking God's guidance" just listed are especially paradigmatic of green. For example, green doesn't, I think, tend to have especially worked-out takes about population ethics. And I, at least, am not saying we *should* take God's input, in all these cases; and still less, to a particular degree. For example, as I've written about previously: I'm not, actually, a big fan of attempts to construe the acts vs. omission distinction in matters-intrinsically (as opposed to matters-pragmatically) terms; I care a lot about possible people in addition to actual people; and I think an adequate ethic of "boundaries" has to move way, way beyond "God created this boundary, therefore it binds."[178]

Nor is God's "input," in any of these cases, especially clear cut. For one thing, God himself doesn't seem especially interested in preventing the extinction of the species he creates. And you're looking for his input re: how to relate to boundaries, you could just as easily draw much bloodier lessons—the sort of lessons that predators and parasites teach. Indeed, does all of eukaryotic life descend from the "enslavement" of bacteria as mitochondria?[179] Or see e.g. this inspiring video (live version here) about "slave-making ants," who raid the colonies of another ant species, capture the baby pupae, and then raise them as laborers in a foreign nest (while also, of course, eating a few along the way). As ever: God is not, actually, a good example; and his Nature brims with original sin.

work. And in some cases, at least, I think there are reasons for pessimism that it can be done at all.

And what if it can't, in a given case? In that case, then the sort of "you must on-reflection have a consistent utility function" vibe associated with Yudkowskian rationality will be even more directly in conflict with taking input-from-God of the relevant kind. Expected-utility-maximizers will *have* to be atheists of that depth. And at a high-level, such conflict seems unsurprising. Yudkowskian rationality is conceives of itself, centrally, as a *force*, a *vector*, a thing that steers the world in a coherent *direction*. But various "input-from-God" vibes tend to implicate a much more constrained and conditional structure: one that asks God more questions (about the default trajectory; about the option set; about existing agents, boundaries, colleges, species, etc), before deciding what it cares about, and how. And even if you *can* re-imagine all of your values from some perspective beyond-the-world—some stance that steps into the void, looks at all possible universe-histories from the outside, and arranges them in a what-I-would-choose-if-I-were-God ranking —still: should you?

[177] Though I think the difference here is somewhat subtle; and both vibes are compatible with the same conclusions.

[178] And re: small-c conservatism: I think that often, if you can actually replace an existing valuable thing with a genuinely-better-thing, you just should. Factoring in, of course, the uncertainties and transition costs and people's-preferences-for-the-existing-thing and all the rest of the standard not-small-c-conservatism considerations. Maybe small-c-conservatism gets *some* weight. But the important question is how much—a question Cohen explicitly eschews.

[179] See this wikipedia for more on the theory. Though obviously, less oppression-vibed narrativizations of this theory are available too.

Queen "slave-maker" (image source here)

Indeed, in some sense, trying to take "guidance from God" seems questionably coherent in the context of your own status as a part of God yourself. That is, if God—as I am using/stretching the term—is just "the Real," then anything you actually do will also have been done-by-God, too, and so will have become His Will. Maybe God chose to create All Souls College; but apparently, if you choose to tear it down, God will have chosen to uncreate it as well. And if your justification for respecting All Souls was that "it's such a survivor"—well, if you tear it down, apparently not. And similarly: why not say that you are *resisting* God, in protecting the Northern White Rhino? The conservation is sure taking a lot of *yang*...

And it's here, as ever, that naturalistic fallacies really start to bite. The problem isn't, really, that Nature's guidance is *bad*—that Nature tells you to enslave and predate and get-your-claws-bloody. Rather, the real problem is that Nature doesn't, actually, give any guidance at all. *Too much stuff* is Nature. Styrofoam and lumber-cutting and those oh-so-naughty sex acts—anything is Nature, if you make it real. And choices are, traditionally, between things-you-can-make-real. So Nature, in its most general conception, seems ill-suited to guiding any genuine choice.

So overall, to the extent green-like respect does tend to "take God's guidance," then at least if we construe the argument for doing so at a sufficiently abstract level, this seem to me like one of the diciest parts of green (though to be clear, I'm happy to debate the specific ethical issues, on their own merits, case-by-case). And I think it's liable, as well, to conflating the sort of respect worth directing at *power* per se (e.g., in the context of game theory, real politik, etc), with the sort of respect worth directing at *legitimate* power; power fused with justice and fairness (even if not, with "my-values-per-se"). I'm hoping to write more about this at some point (though probably not in this series).

That said, to the extent that deep atheism takes the *general* naturalistic fallacy—that is, the rejection of any move from "is" to "ought"—as some kind of trump-card objection to "taking guidance from God," and thus to green, I do want to give at least one other note in green's defense: namely, that insofar as it wishes to have any ethics at all, many forms of deep atheism need to grapple with some version of the general naturalistic fallacy as

well.

In particular: deep atheists are ultimately *naturalists*. That is, they think that Nature is, in some sense, the whole deal. And in the context of such a metaphysics, a straightforward application of the most general naturalistic fallacy seems to leave the "ought" with nowhere to, like, attach. Anything *real* is an "is"—so where does the "ought" come from? Moral realists love (and fear) this question—it's their own trump card, and their own existential anxiety. Indeed, along this dimension, at least, the moral realists are even more non-green than the Yudkowskians. For unlike the moral realists, who attempt (unsuccessfully) to untether their ethics from Nature entirely, the Yudkowskians, ultimately, need to find some ethical foothold *within* Nature; some bit of God that they *do* take guidance from. I've been calling this bit your "true self," or your "heart"—but from a metaphysical perspective, it's still God, still Nature, and so still equally subject to whatever demand-for-justification the conceptual gap between *is* and *ought* seems to create.[180] Indeed, especially insofar as straw-Yudkowskian-ism seems to assume, specifically, that its true heart is closely related to what it "resonates with" (whether emotionally or mentally), those worried about naturalistic fallacies should be feeling quite ready to ask, with Lewis: why *that?* Why trust "resonance," ethically? If God made your resonances, aren't you, for all your atheism, taking his guidance?[181]

Indeed, for all of the aesthetic trappings of high-modernist science that straw-Yudkowskianism draws on, its ethical vibe often ends up strangely Aristotelian and teleological. You may not be trying to act in line with Nature as a whole. But you are trying to act in line with *your* (idealized) Nature; to find and live the self that, in some sense, you are "supposed to" be; the true tree, hidden in the acorn. But it's tempting to wonder: what kind of naturalistic-fallacy bullshit is that? Come now: you don't have a Nature, or a Real Self, or a True Name. You are a blurry jumble of empirical patterns coughed into the world by a dead-eyed universe. No platonic form structures and judges you from beyond the world—or least, none with any kind of intrinsic or privileged authority. And the haphazard teleology we inherit from evolution is just that. You who seek your true heart—what, really, are you seeking? And what are you expecting to find?

I've written, elsewhere, about my answer—and I'll say a bit more in my next essay, "On attunement," as well. Here, the thing I want to note is just that once you see that (non-nihilist) deep atheists have naturalistic-fallacy problems, too, one might become less inclined to immediately jump on green for running into these problems as well. Of course, green often runs into much more specific naturalistic-fallacy problems, too—related, not just to moving from an *is* to an *ought* in general, but to trying to get "ought" specifically from some conception of what Nature as a whole "wants." And here, I admit, I have less sympathy. But all of us, ultimately, are treating some parts of God as too-be-trusted. It's just that green, often, trusts more.

[180] Per standard meta-ethical debates, I'm counting abstracta as parts of Nature and God, insofar as they, too, are a kind of "is." I think this maybe introduces some differences relative to requiring that anything Natural be concrete/actual, but I'm going to pass over that for now.

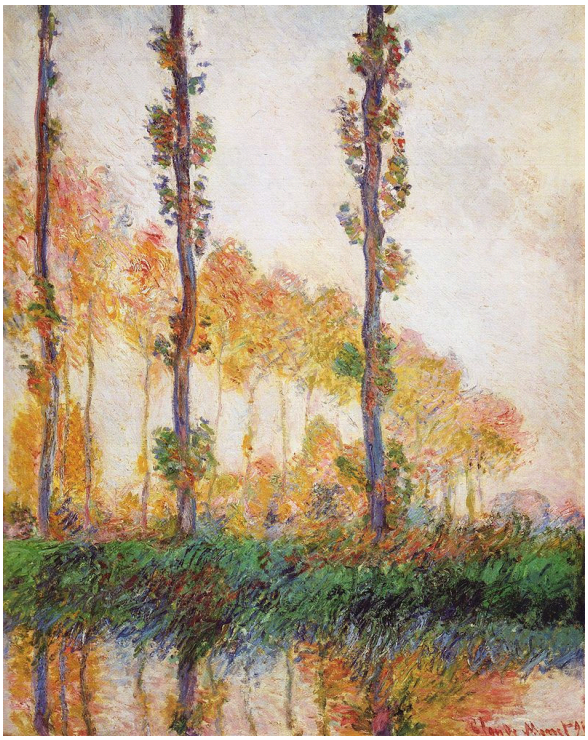
[181] Well, we should careful. In particular: your resonances don't need to be resonating *with themselves*—rather, they can be resonating with something else; something the actual world, perhaps, never dreamed of. But if you later treat *the fact that you resonated with something* as itself ethically authoritative, you are giving your resonances some kind of indirect authority as well (though: you could view that authority as rooted in the thing-resonated-with, rather than in God's-having-created-the-resonances).

Chapter IX

On attunement

"You, moon, You, Aleksander, fire of cedar logs.
Waters close over us, a name lasts but an instant.
Not important whether the generations hold us in memory.
Great was that chase with the hounds for the unattainable meaning of the world."

—Czeslaw Milosz, "Winter"



"Poplars (Autumn)," by Claude Monet (image source here)

The last chapter examined a philosophical vibe that I (following others) call "green." Green is one of the five colors on the Magic the Gathering Color Wheel, which I've found (despite not playing Magic myself) an interesting way of classifying the sort of the energies that tend to animate people.[182] The colors, and their corresponding shticks-according-to-Joe, are:

- *White*: Morality.
- *Blue*: Knowledge.

[182]My relationship to the MtG Color Wheel is mostly via somewhat-reinterpreting Duncan Sabien's presentation here, who credits Mark Rosewater for a lot of his understanding. What I say here won't necessarily resonate with people who actually play Magic.

- *Black*: Power.
- *Red*: Passion.
- *Green*: ...

I haven't found a single word that I think captures green. Associations include: environmentalism, tradition, spirituality, hippies, stereotypes of Native Americans, Yoda, humility, wholesomeness, health, and *yin*. My last chapter tried to bring the vibe that underlies these associations into clearer view, and to point at some ways that attempts by *other colors* to reconstruct green can miss parts of it. In particular, I focused on the way green cares about *respect*, in a sense that goes beyond "not trampling on the rights/interests of moral patients" (what I called "green-according-to-white"); and on the way green takes *joy* in (certain kinds of) *yin*, in a sense that contrasts with merely "accepting things you're too weak to change" (what I called "green-according-to-black").

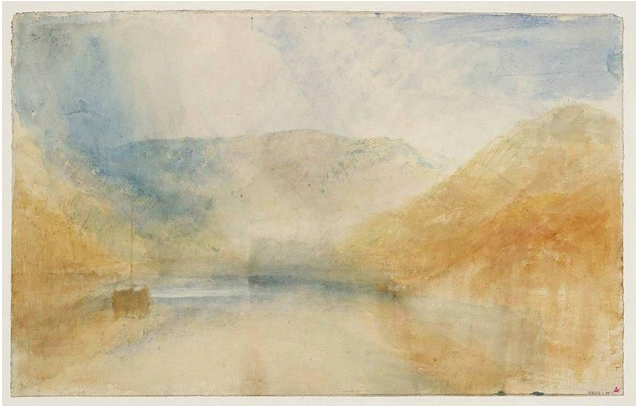
In this chapter, I want to turn to what is perhaps the most common and most compelling-to-me attempt by another color to reconstruct green—namely, "green-according-to-blue." On this story, green is about making sure that you don't act out of inadequate *knowledge*. Thus, for example: maybe you're upset about wild animal suffering. But green cautions you: if you try to remake that ecosystem to improve the lives of wild animals, you are at serious risk of not knowing-what-you're-doing. And see, also, the discourse about "Chesterton's fence," which attempts to justify deference towards tradition and the status quo via the sort of knowledge they might embody.

I think humility in the face of the limits of our knowledge is, indeed, a big part of what's going on with green. But I think green cares about *having* certain kinds of knowledge too. But I think that the type of knowledge green cares about *most* isn't quite the same as the sort of knowledge most paradigmatically associated with blue. Let me say more about what I mean.

1 How do you know what matters?

"I went out to see what I could see ..."

—Annie Dillard, "Pilgrim at Tinker Creek"



An 1828 watercolor of Tintern Abbey, by J.M.W. Turner (Image source here.)

Blue, to me, most directly connotes knowledge in the sense of: science, "rationality," and making accurate predictions about the world. And there is a grand tradition of contrasting this sort of knowledge with various other types that seem less "heady" and "cognitive"—even without a clear sense of what exactly the contrast consists in. People talk, for example, about intuition; about system 1; about knowledge that lives in your gut and your body; about knowing "how" to do things (e.g. ride a bike); about more paradigmatically social/emotional forms of intelligence, and so on.

And here, of course, the rationalists protest at the idea that rationality does not encompass such virtues (see, e.g., the discourse about "Straw Vulcans"). Indeed, if we understand "rationality" as a combination of "making accurate predictions" (e.g. "epistemic" rationality; cf blue) and "achieving your goals" (e.g., "instrumental" rationality; cf black), then an extremely broad variety of failures—e.g., social/emotional clumsiness, indecision, overthinking, disconnection from your intuition, falling-off-your-bike—can count as failures of rationality. With blue and black accounted for, then, is anything left over?

Well, yes—especially if we're thinking of rationality as Yudkowsky does, in the context of the sort of meta-ethical anti-realism I discussed in "Deep atheism and AI risk." In particular: I've written, previously, about the sense in which anti-realist rationality stumbles in the realm of ethics and value.

> "Give anti-realist rationality a goal, and it will roar into life. Ask it what goals to pursue, and it gets confused. 'Whatever goal would promote your goals to pursue?' No, no, that's not it at all."

Or put another way: anti-realist rationality has a very rich concept of "instrumental rationality," but a very impoverished concept of what we might call "terminal rationality"—

that is, of how to do the "what matters intrinsically?" thing right. It tells you, at least, to not fail on the blue-and-black thing—to not form terminal goals based on a mistaken or incomplete picture of the world, or of what-will-lead-to-what. But beyond that, it goes silent.

Where, then, do your terminal goals come from? Well, for the most standard form of anti-realist rationality, from *red*. That is, from your heart, your desire, your passion—Hume's famous slavemaster. That is, for all its associations with blue (and to a lesser extent, black), rationality (according to Yudkowsky) is actually, ultimately, a project of *red*. The explanatory structure is really: red (that is, your desires), *therefore* black (that is, realizing your desires), *therefore* blue (knowledge being useful for this purpose; knowledge as a form of power). Blue is twice secondary—a tool for black, which is itself a tool for red. (Of course, red can also value blue for its own sake—and perhaps this ultimately a better diagnosis of what's going on with many rationalists. But from a philosophical perspective, intrinsically valuing knowledge is much more contingent.)

Indeed, in this sense, it's not just green that anti-realist rationality struggles to capture. It's also *white*—that is, morality. Anti-realist rationality has a concept of *cooperation*, in the sense of "getting-to-the-Pareto-frontier," "making trade agreements," and so on (with various fancy decision theories potentially playing a role in the process). But as I've written about previously, this sort of cooperation is too much a project of power to really capture morality—and in particular, it's much too willing to kill, lie, defect, etc in interactions with weaker, dumber, and/or unloved-by-the-powerful agents (this is core to why Yudkowsky doesn't expect the AIs, for all their black-and-blue, to be nice to humans).[183]

And beyond this type of cooperation, what sort of white is left for anti-realist rationality? Just: whatever sort of white you happen to be red about. That is: morality is just one possible thing-your-heart-could-care-about, among many others. It's another brand of paperclips. Should we have a color for paperclips as well? And for staples? And for staples-of-a-slightly-different-shape? And morality, too, comes in many different shapes. Which morality do we mean?

Indeed, for all the social connections between the Yudkowskian rationalists and the effective altruists, the philosophical connection, here, starts to break down. Effective altruism, as a philosophical project, tends to assume that there is this thing, "goodness," which EAs try to maximize; or this thing, "altruism," which EAs try to do effectively.[184] But Yudkowskian rationalism doesn't, actually, have a privileged concept of "goodness," or of "altruism" (see my essay "In search of benevolence" for more on this). Rather, there are a zillion concepts in the broad vicinity, which different hearts can latch onto differently—and it's not clear what distinguishes them, deeply, from other sorts of goals or hobbies.

No wonder, then, that many of the philosophical founders of effective altruism (e.g. Singer, Parfit, Ord, MacAskill) tend towards moral realism. Effective Altruism is a lot about Morality with a capital M. Maybe it presents itself, in various contexts, as just-another-hobby. And sure, hobbyists are welcome. But various strands of philosophical EA want, underneath, to act with the righteousness of a True Church—to be doing, you know, the

[183] See Soares here for more.

[184] Within constraints; with the part of their lives they choose to devote to this activity; etc ...

Good Thing, the Right Thing; and to be doing it the best way; the way you, like, *should*. Maybe you're not *obligated* to do this (rather, it's "supererogatory.") And sure, you're too *weak* to do it fully. But God smiles brighter as you do it more.

And this self-conception fits uncomfortably with treating white as ultimately grounded in red; morality as ultimately grounded in passion or sentiment. White wants *God's* heart to smile on it; its own heart is beside the point, and lacks the authority white seeks.[185] That kind of authority, thinks paradigmatic white, needs to be more objective. It needs to speak with the world's voice—a voice that says to the reflectively-coherent suffering-maximizers "you are *wrong*" and not just "you and I want different things, and I'm ready to fight about it." And where does one go to call other people wrong? Standardly: to *blue*. That is, paradigmatic capital-M Morality wants its shtick to follow from (and be a form of) knowledge. Blue-therefore-white.[186] But anti-realism about meta-ethics denies morality this objectivity. Morality seeks grounding in blue; but red is the best it can get.

Right? Well, at some level: yes, probably. But I worry about telling the story too crudely, and in the wrong order. In particular: I worry that trying to ground ethics in either paradigmatically blue-style knowledge, or paradigmatically red-style passion, or in some combination, misses some other, more elusive dimension of normative epistemology—something neither paradigmatically red nor blue (even if, ultimately, it can be built out of red-and-blue); and something closely associated with wisdom. I'll call this dimension "attunement."

2 Gestures at attunement

> *"Don't look upon the light in your eyes, look upon the sky.*
> *And don't feel the pain in your side, feel the wound there ...*
> *Don't hear my words, hear the roughness and warmth of my mind.*
> *Meet me here, face to face."*
>
> *—Katja Grace, "As you know yourself"*

What is attunement? I'm thinking of it, roughly, as a kind of meaning-laden receptivity to the world.[187] Something self-related goes quieter, and recedes into the background; something beyond-self comes to the fore. There is a kind of turning outwards, a kind of openness; and also, a kind of presence, a being *in* the world. And that world, or some part of it, comes forward as it always has been—except, often, strangely new, and shining with meaning.

Here's a passage from Marilyn Robinson's "Housekeeping" that evokes attunement for me:[188]

[185] This lack of authority is key to the intuitive pull of Lewis's position, in the *Abolition of Man*, that anti-realists influencing the values of others must be tyrants.

[186] And then, for EA, therefore-black-therefore-blue-again.

[187] Webster's dictionary for "attune" says: "(1) to bring into harmony, (2) to make aware or responsive."

[188] Indeed, I think part of what's compelling about Robinson's writing is her degree of attunement; the way the world, in her vision, seems to shine with quiet holiness; the way plain things appear somehow numinous. I associate Virginia Woolf with some quality in this vicinity, too—though of a different flavor than Robinson.

> What was it like. One evening one summer she went out to the garden. The earth in the rows was light and soft as cinders, pale clay yellow, and the trees and plants were ripe, ordinary green and full of comfortable rustlings. And above the pale earth and bright trees the sky was the dark blue of ashes. As she knelt in the rows she heard the hollyhocks thump against the shed wall. She felt the hair lifted from her neck by a swift, watery wind, and she saw the trees fill with wind and heard their trunks creak like masts. She burrowed her hand under a potato plant and felt gingerly for the new potatoes in their dry net of roots, smooth as eggs. She put them in her apron and walked back to the house thinking, What have I seen, what have I seen. The earth and the sky and the garden, not as they always are. And she saw her daughters' faces not as they always were, or as other people's were, and she was quiet and aloof and watchful, not to startle the strangeness away.

Zadie Smith writes about another example. For much of her life, she hated the music of Joni Mitchell. It just sounded like noise: "a piercing sound, a sort of wailing." Then, one day, she was visiting Tintern Abbey with her husband. He had Joni on in the background in the car. Smith hated it as always. They parked.

> "I opened a car door onto the vast silence of a valley. I may not have had ears, but I had eyes. I wandered inside, which is outside, which is inside. I stood at the east window, feet on the green grass, eyes to the green hills, not contained by a non-building that has lost all its carved defenses ... And then what? As I remember it, sun flooded the area; my husband quoted a line from one of the Lucy poems; I began humming a strange piece of music. Something had happened to me..."

Tintern Abbey (Image source here)

Exactly what happened isn't clear. But Smith's experience of Joni Mitchell changes dramatically:

> How is it possible to hate something so completely and then suddenly love it so unreasonably? How does such a change occur? ... This is the effect that listening to Joni Mitchell has on me these days: uncontrollable tears. An emotional overcoming, disconcertingly distant from happiness, more like joy—if joy is the recognition of an almost intolerable beauty. It's not a very civilized emotion.

Smith's essay emphasizes the *yin* at stake in the attunement[189]—the listening, the letting-in—and also, the sense of recognizing something intensely (intolerably?) important, to

[189] "I can't listen to Joni Mitchell in a room with other people, or on an iPod, walking the streets. Too risky.

which it is possible to be blind, or inadequately sensitive.[190] I've written about this before: "seeing more deeply," "the doorway to real life." I think experiences of beauty, spirituality, morality, and meaning all often involve a sense of attunement in this sense. And I think green cares a lot about that.

Indeed, what is Ogion trying to teach Ged, in silence, in the eyes of animals, and the flights of birds? *The Wizard of Earthsea* talks a lot about "true names"—but how do you learn them? Foster, in My Octopus Teacher, is trying to learn. And I think green-like figures of wisdom—Yoda, the Buddha, the archetype of an "elder"—often have very strong attunement vibes.

Admittedly, I'm painting in fairly broad strokes here. But hopefully, for present purposes, it's enough of a gesture.

3 Attunement and your true heart

"You, music of my late years, I am called
By a sound and a color which are more and more perfect.

Do not die out, fire. Enter my dreams, love.
Be young forever, seasons of the earth."

—Czeslaw Milosz, "Winter"

"Hunters in the snow," by Pieter Bruegel the Elder (image source here)

I can never guarantee that I'm going to be able to get through the song without being made transparent—to anybody and everything, to the whole world. A mortifying sense of porousness. Although it's comforting to learn that the feeling I have listening to these songs is the same feeling the artist had while creating them: 'At that period of my life, I had no personal defenses. I felt like a cellophane wrapper on a pack of cigarettes."

[190] Though I sympathize somewhat with Katja Grace, who finds that Smith's essay as a whole doesn't quite say what she wanted it to say.

> "I don't think the words meant what I wanted them to mean, but it was arguably about what I wanted it to be about, and left me with the message I wanted. Which I somehow believe might be what she meant to mean, especially now that I try to find my own words. It sounded like she was saying,"if you lower your boundaries and give time to various initially unappealing art forms, they can be awesome". But that's a message in the wrong register. What I wanted it to say was, open yourself in some deep way, turn yourself around, open eyes that you didn't know you had, and everything might touch you. Touch you like you are its edges and its texture and you know everything, even if you can't put it into words—not just some heightened tendency to mindless tears, or another 'positive mental state' for the utility logs. Don't ask for more reasons on your blind and empty abstracta table, be your soul instead, and press yourself against the world, into the world. Hear every cell itself, not the trace it leaves in your proposition set. 'Attunement."'

Now: when I wrote about attunement previously, under the heading of "seeing more deeply," I said that it tends to pull me towards realism about value. This is centrally because it seems like it discloses something simultaneously beyond-myself *and* valuable/important. That is, it has all the *yin* of blue—of knowledge, of *receiving*. But the thing-received, the thing-known, is something normative and meaningful.

Indeed, experiences of attunement are core to my own moral epistemology, and to my spirituality more generally. Philosophy, sure. But ultimately, for so many of us, it's our deepest experiences that lead us onward. Some vision, some seeing, that says "this, this; don't forget." And said in some distinctive way; not as just-another-emotion, but with, it seems, some different depth—some particular harmony and clarity. For me, at least, this sort of depth is core to the weight and mystery and authority of that strange word, "goodness." It's related, I think, to the way sincerity feels like coming home; like something falling into its proper place. Chögyam Trungpa talks about "basic sanity."

Does meta-ethical anti-realism preclude blue from receiving words like "goodness"? Blue alone: yes. And indeed, I expect that attunement will ultimately be a matter of both blue and red: of knowledge and love, your eyes and your heart, intermixed. But how do you see with your true heart's eyes? Blue's most paradigmatic answer is: "learn the facts; get 'full information.' " But that doesn't seem like it captures what's going on with attunement very directly. In particular: experiences of attunement often feel much more like "realization" than like a change in belief. It's often the same old facts; but with new resonance, new intensity, a new remembering.

And if we ask paradigmatic red to identify your true heart, it's not clear that we capture attunement very well, either. In particular: paradigmatic red calls to mind a tumble of different passions and desires, colliding with each other in a contest of raw power—and the king of the hill gets to be Hume's slavemaster.[191] Is your true heart, then, simply the strongest contestant, or coalition?[192] But at the least, insofar as the thing disclosed via attunement claims to be your true heart, it does not do so on the basis of felt intensity—or at least, not only. Hunger and lust, pride and fear—these can easily be more intense, at some level, than experiences of attunement. And they are quite a bit more common; quite a bit *easier*. Are they not, then, the truer red? Yet amidst all the shouting of ten thousand often-louder voices, when attunement speaks, the room goes quiet. And when attunement leaves, the room tries, so hard, to remember what it said, and to call it back again.

Of course, we can try to construct a story about your "true heart" that captures this dynamic. "I just do trust some experiences of care more than others. They just do leave a deeper and more sustained mark on my motivations and my orientation towards the world; and this is what makes them my true heart." And ultimately, maybe something like this is the right story. That is, perhaps, for those of us for whom something like "attunement" plays a key role in shaping our core values (I don't think this is everyone), this itself is centrally a fact about our particular pattern of care and meta-care; about how we do red.

[191] This is the psychological microcosm of Yudkowsky's cosmic narrative.
[192] Plus, presumably, a bunch of other idealization?

If so, though, it seems like a very important fact to understand. Apparently, I trust certain types of experiences/ways-of-being vastly more than others to shape what I do with my one and only life. Apparently, some experiences/ways-of-being disclose something that is, to me, searingly and intolerably important. And this sort of experience seems to be associated, most centrally, not with paradigmatic red, or with paradigmatic blue, but with green—whatever *that* is.

4 Green, therefore ...

"Quit your tents."

—Annie Dillard, "Teaching a Stone to Talk"

"Moses on Mount Sinai" by Jean-Léon Gérôme (image source here)

And even if green can/should ultimately be built out of red and blue, we should make sure to tell the story in the right order. Here I think of a friend of mine, who identifies very strongly with morality, and with Effective Altruism. I told him my theory that paradigmatic Effective Altruism wants the story to be: knowledge, therefore morality. He said that for him, it feels like the story is more like: morality, therefore knowledge-therefore-morality. Or perhaps more accurately: morality, therefore: whatever it is such that therefore-morality. That is: the primary allegiance is to morality, *whatever that is; whatever grounds it*.[193] He is moralist, first; and meta-ethicist, second.[194]

I think something similar might be true for me (or parts of me), except with green first, instead—and in particular, green qua spirituality, green qua attunement. That is, I think the core story for me may be: green, therefore: whoa, that was important. How do I

[193] I have another friend, who also identifies very strongly with Morality, who thinks I shouldn't be allowed to say that white's shtick is "morality," because all the other colors presumably think that their shtick is The Moral Thing, too. But I think she is wrong about what the other colors think.

[194] Or maybe: not at all? Can we just ignore meta-ethics, please? Isn't it a bit of a verbal dispute?

honor and do right by whatever that was? How do I see and respond to whatever I just saw-in-part? The earth and the sky and the garden. "The real world." But how do you live there?

I do a lot of morality stuff. But a lot of it feels like green-therefore-white; morality as a way of honoring and responding to whatever-green-saw.

Of course, I do a lot of meta-ethics, too. I try to see green, too, more whole, and to figure out the right therefores—the true role of red, and of blue; the true nature of white. But the map is not the territory; I am more than my theory of myself; and my allegiance to seeing and responding to the world seen-by-attunement outstrips my confidence in any particular story about what grounds this allegiance, or of what-attunement-sees. This isn't to say that blue can't alter my attitude towards green (and green-without-blue isn't real attunement, anyway). And no part of us needs to be the ultimate foundation—each can build and support and critique the others. But green, for me, is first and foremost according-to-itself: beauty, holiness, grace—raw and unrationalized.

Here I think of Robinson again:

> Something happened, something so memorable that when I think back to the crossing of the bridge, one moment bulges like the belly of a lens and all the others are at the peripheries and diminished. Was it only that the wind rose suddenly, so that we had to cower and lean against it like blind women groping their way along a wall? or did we really hear some sound too loud to be heard, some word so true we did not understand it, but merely felt it pour through our nerves like darkness or water?

Too loud to be heard. How do you know it's true if you can't hear, or understand? But I think we should keep listening—keep taking attunement on its own terms—regardless. This is partly because I think we *do* understand the thing-attunement-sees, at some level—that's why, I think, that goodness and beauty and holiness feel so much like coming home. "I had been my whole life a bell," writes Annie Dillard, "and never knew it until at that moment I was lifted and struck."[195] And it's often thought that if something makes you-ring-like-a-bell, this is connected with stuff about your true heart. Hence the word "resonance."

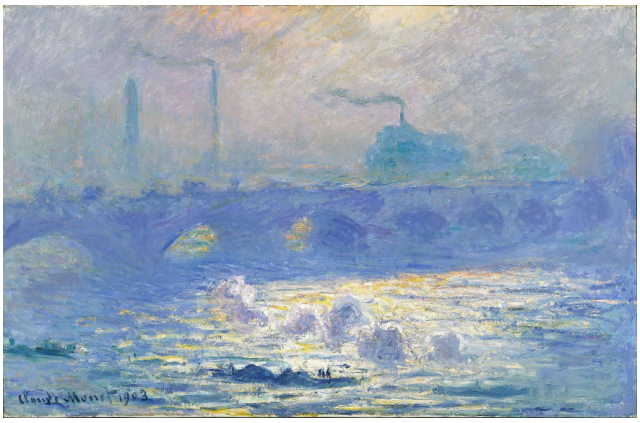
But also, to the extent we don't yet have a settled story about what's-up-with-attunement; to the extent blue does not yet grok green; still, I think that's OK, too and that we should keep doing green-therefore-green in the meantime. In particular: when it comes to seeing with our true heart's eyes, I think we should acknowledge how much we are still, yet, as blind women, groping our way. Anti-realists, at least, don't actually have a clear story about their true hearts; and the realists don't have the eyes they need to see God's heart, either. But we still need to cross the bridge; and it is still extremely possible to fall. Indeed: the invention of AGI is a very big bridge. Maybe the biggest. And we might all fall at once.

It's similar to what happens when we talk about "wisdom." I say, often, that I want the future to be "wise." But what does that mean? Again, doing blue-and-black right

[195] From *Pilgrim at Tinker Creek*.

is a start, but it's not enough; you have to get the ethics and meaning thing right, too. And how do we do that? We don't know, we don't know. We have scattered glimpses—histories; mistakes; lessons-learned. We have logic and empathy and imagination. We have traditions, archetypes, stories. And we have attunement. But we don't have a settled program for becoming-wise. And we need to do it anyway.

To the extent green is the "wisdom" color, then, I think we should be pretty interested in making sure we're staying in touch with green-on-its-own-terms. And indeed, when I think about the sort of wise that I want future people to be, I imagine them having the attunement thing in spades—some kind of intensity and tenderness and vastness of consciousness, some deep receptivity and responsiveness. If what one learns, from attunement, is "basic sanity," I want the future to be *sane*.



"Waterloo Bridge," by Claude Monet (Image source here)

5 A future without attunement

"Who is there to carry on the life-thread of Wisdom?"

—Hakuin

And how fragile is sanity? I have a friend, a moral realist,[196] who worries that the sort of Yudkowskian anti-realism pervasive amongst AI folks will create the world in its own image. That is, this sort of anti-realism assumes that the only type of agent you can build is a generalized paperclipper, a "Hume-bot," chugging away in pursuit of its arbitrary preferences, instead of turning outwards towards the world, and seeking after some truer and deeper vision of meaning and morality. Aren't we all Hume-bots, after all? That's how the anti-realists model themselves, at least. So, worries my friend, the conditioners will make future agents—human and artificial—in their own self-image; blind, not just to the content of the *Tao*, but to the *existence* of the *Tao*; asking only, ever, about what they want, rather than about what's right, what's good, what's worthy. And thus, the True Way will be lost forever; and the world will go blind.

[196] Or at least, someone quite a bit more sympathetic to moral realism than me.

I think my friend is too confident about moral realism (and/or, too willing to wager on it). But I think he's pointing at a real concern—and I think it's a concern that anti-realists can share. To put it in my own terms: I think it's a concern about a future that has lost attunement. Whatever our meta-ethics, we can agree that there is a thing that humans do, when they turn outwards, and with their hearts open; seeking, in *yin*, some truer contact with the good; trying to listen more deeply to the great song of the world. And we can agree that this thing, and the thing-it-discloses when done well, is profoundly precious. We want a future where it flowers fully; a future that sees in full, and with our whole hearts, what we now see only in part. Maybe the true story about this looks more like moral realism, or moral anti-realism; or, perhaps more likely, like neither in its current self-conception. But regardless, we want the future to cross the bridge, and with its soul intact. To finish, or to follow ever deeper, that most ancient pilgrimage: from cave to sun; from dream to the vast and waking world.

"Who is there to carry on the life-thread of Wisdom?" writes Hakuin. Who indeed. But if red-without-green grows unwise, then a future that runs only on red, or on red-therefore-black-therefore-blue, might lose the life-thread. Or to put it in more familiar Yudkowskian ontology: to the extent that whatever is going on with green, and with attunement, is itself core to our real red, our true hearts, then a future without attunement has made its heart false.

Indeed, I think we can read Lewis, in the Abolition of Man, as worried about something similar.[197] He wishes for a regenerate science such that, "while studying the *It*, it would not lose what Martin Buber calls the *Thou*-situation." And attunement, to my mind, is closely related to approaching the world as a *Thou*—to that particular sort of *yin* that seeks, not just knowledge, but *encounter*; to give the world, the Other, its own dignity; to feel the weight of its being; be present *with* something else that is present, too. "Don't look upon the light in your eyes, look upon the sky," writes Katja Grace. "Meet me here, face to face."

Of course, Lewis presents his concern, centrally, as about whether we will stay "within the *Tao*." But if we think of the *Tao* less as The Objectively True Morality that All Cultures Have Basically Agreed On, and something more like "life lived from attunement," then I start to feel better about passages like the following:

> In the *Tao* itself, as long as we remain within it, we find the concrete reality in which to participate is to be truly human: the real common will and common reason of humanity, alive, and growing like a tree, and branching out, as the situation varies, into ever new beauties and dignities of application.

I do think that attunement participates in some concrete reality—something that draws us more deeply into our humanity, and into what I've called "real life." And reframed in such terms, I think this passage actually gets at something pretty core; something that I very much want the age of AGI to stay "within," and for anyone remotely nearby the power of a "conditioner" to remain especially in-contact-with. Indeed, it sounds a lot

[197] Like my friend above, Lewis focuses on realism vs. anti-realism about meta-ethics, but I don't think we need to follow him in this.

like bits of Yudkowsky's own poetry about Coherent Extrapolated Volition. If we had grown up farther together. If we were more the people we wished we were. And when we imagine the path to good futures, I, at least, do actually imagine something akin to a civilization "alive and growing like a tree"—the way we've already been growing, painfully, over the centuries. A process that consists, centrally, not in the conditioners "figuring out the right values" and then "executing," but rather in some kind of organic and ongoing self-adjustment; the way a plant grows, gradually, towards the light.[198]

"The Old Oak," by Jules Dupre (image source here).

[198] From Christiano (2021):

"I expect technology could radically transform the world on a timescale that would be disorienting to people, but for the most part that's not how we *want* our lives to go in order to have the best chance of reaching the best conclusions about what to do in the long run. We do want some effects of technology—we would like to stop being so hungry and sick, to have a little bit less reason to be at each other's throats, and so on—but we also want to be isolated from the incomprehensible, and to make some changes slowly and carefully.

So I expect there to be a very recognizable thread running through humanity's story, where many of the humans alive today just continue to being human and growing in a way that is familiar and comfortable, perhaps changing more quickly than we have in the past but never so quickly that we are at risk of losing our footing. The point of this is not because that's how to have the best life (which may well involve incomprehensible mind-alteration or hyper-optimized virtual reality or whatever). It's because we still have a job to do.

The fact that you are able to modify a human to be much smarter does not mean that you need to, and indeed I think it's important that you take that process slow. The kinds of moral change we are most familiar with and trust involve a bunch of people thinking and talking, gradually refining their norms and making small changes to their nature, raising new generations one after another.

...I think that the community of humans taking things slowly and living recognizable lives isn't an irrelevant sideshow that anyone serious would ignore in favor of thinking about the crazy stuff AI is doing "out there" (or the hyper-optimized experiences some of our descendants may immerse themselves in). I think there's a real sense in which it's the main thread of the human story; it's the thread that determines our future and gradually expands to fill the universe."

6 Primal blue

"Go where those others went to the dark boundary
for the golden fleece of nothingness your last prize

go upright among those who are on their knees
among those with their backs turned and those toppled in the dust

you were saved not in order to live
you have little time you must give testimony"

—Zbigniew Herbert, "The Envoy of Mr. Cogito"

"Jason and Eros" by Gustave Moreau (image source here)

So far I've been talking about attunement centrally in terms of the normative stuff that it discloses. But various aspects of attunement also seem associated with non-normative types of knowledge—with a familiar sort of blue. For example: perception. To look upon the sky, rather than the light in your eyes, means to retain your grip not just on the raw data the perception provides, but on the *function* of perception—namely, to *refer;* to make

contact; to see past the light to the thing-shining; to carve the right meaning from the noise. And whatever their other spiritual and normative connotations, ways of being in the vicinity of being present/mindful/"awake" seem to be doing something pretty blue as well—something directly related to the mundane (or at least, non-normative) truth.

That said: is it the same sort of mundane truth at stake when you make predictions, or improve your model of the world? I'm not sure. Certainly, failing to be "present" can easily lead to prediction-problems. But internally, various sorts of "presence" and "awakeness" often feel less propositional, and more like "getting a grip." Like the same-old world coming into focus. "Poise." It's a type of blue related to that particular and especially-strange sort of knowledge that consciousness can have of itself—the thing that happens when, let's just check one more time: yep, not-a-p-zombie.

Is that so basic? Descartes thought so. But the illusionists disagree. Regardless, though: the question burns hot and unanswered: *what's going on when your consciousness "knows itself"?* Or relatedly, when you know that you exist, or that anything exists, or that you are some particular being among others? What sort of world looks out of *any* eyes—let alone, so many different and apparently-separate sets at once? Meet me *here*—but where is that?

Basic sanity is basic partly in the intimacy of its contact with this sort of primordial ground—Reality, being-a-thing, being-in-the-world, being-aware-at-all. It remains, at least to me, quite a mystery. In this sense, I associate attunement, not just with wisdom-qua-ethics, but with various other sorts of brute and not-understood-by-me sorts of existential awareness—what we might call green-blue; or maybe better, "primal blue." Primal because I think there is something raw and animal-like about the way we know stuff like "I'm conscious" and "the world exists." We know it before we know-what-we're-knowing. We know it using the very foundations of our minds. No wonder, then, that it's not going to win any prediction-markets—the foundations are priced in everywhere. But turn, directly, towards the foundations themselves, and they become a coal-face, and you start to touch raw rock.

Whether in ethics or elsewhere, I think that attunement is partly about this sort of living-at-the-coal-face; placing your mind, fully and openly, at its own edges; letting it propel itself towards the Real—that most-here, most-beyond—with its whole energy. And the coal face requires awareness that can exceed understanding; the ability to make *contact* with something not-mastered, not-understood. Some of this is about having the sort of map that lets the territory speak—that classic virtue of basic perception. But I think there's also something else, related to being alive first, and making maps second, and in service. And of going, when necessary, without maps, to that dark boundary, where the others went. "Be your soul," writes Katja Grace. "Press yourself against the world, into the world."

7 Being your soul

"Beware of dryness of heart love the morning spring
the bird with an unknown name the winter oak

light on a wall the splendour of the sky
they don't need your warm breath
they are there to say: no one will console you

be vigilant—when the light on the mountains gives the sign—arise and go
as long as blood turns in the breast your dark star

repeat old incantations of humanity fables and legends
because this is how you will attain the good you will not attain
repeat great words repeat them stubbornly
like those crossing the desert who perished in the sand"

—Zbigniew Herbert, "The Envoy of Mr. Cogito"

"The Red Hills, Lake George," by Georgia O'Keeffe (Image source here)

And I think there is a connection between *being* our souls, in this sense, and keeping them intact as we cross the bridge into the age of AGI. Consider Lewis's worry at the end of *Abolition of Man*.

> The whole point of seeing through something is to see something through it. It is good that the window should be transparent, because the street or garden beyond it is opaque. How if you saw through the garden too? It's no use trying to "see through" first principles. If you see through everything, then everything is transparent. But a wholly transparent world is an invisible world. To "see through" all things is the same as not to see.

What is this bit really about? Most of the book is about ethics—but the "first principles," here, aren't necessarily ethical. And indeed, I think Lewis is actually gesturing at (though: not explaining) a broader argument, made in e.g. Miracles, that Reason can't view itself

as a purely natural process, because (roughly) it is trying to follow universal laws of Reason, which is too deeply different from being subject to brute causation.[199] That is, Lewis thinks naturalism will not just kill the *Tao*; it will kill logic and math and all the other sorts of sanity (and hence, in his view, naturalism is insane). My impression is that various strands of both analytic and continental philosophy have similar worries about truth, reference, and other epistemic basics—that in becoming pure-causation, the world dissolves into a play of pure power. It's not just that we lose contact with Objective Morality; we lose, also, the "contact" of truth and reference and encounter; and the only type of contact that remains is collision.

I'm not, generally, a fan of various arguments in this vicinity. Re: Lewis's for example: I think that brute causal processes can themselves follow universal laws of Reason—see, e.g., a theorem-prover, or this marble adding machine—and that Reason can itself develop an empirical and naturalistic world-picture that validates that this is what's happening with parts of our brain. And I suspect that we will be able to give similarly self-validating stories about stuff like reference as well, though I've thought less about the topic.

But I also think that Lewis's argument, at least, is still pointing at something interesting re: "seeing through first principles"—and something related to being what Yudkowsky calls being "created already in motion."[200] In particular: even though Reason is a brute causal process (though: not *merely* one), and can come to see (and validate) itself as such, the process of so-seeing requires *being* Reason—*being* your soul—as opposed to merely modeling it. You do, in fact, have to stay *within* something. You have to *think*—to seek, yourself, whatever it is that Reason seeks; to *be* the onrush of that part of your being. And this requires what I've previously called "living from the inside," and "looking out of your own eyes," instead of only from above. In that mode, your soul is, indeed, its own first principle; what Thomas Nagel calls the "Last Word." Not the seen-through, but the seer (even if also: the seen).

Lewis seems to think that naturalists can't do this (or: not consistently). That naturalists, by being too much within-the-world, have been somehow cast forever outside themselves. As I discussed in a previous essay, I think he's wrong—and that his mistake, here, is closely related to why he seems to wrongly assume that naturalists must lose their grip on any non-crass ethics. Just as naturalists can *be* their reason and their logic, they can *be*

[199] See this Wikipedia for more on this argument's development in analytic philosophy, and see also Thomas Nagel's The Last Word, which I think develops a somewhat similar line of thought. Here's Lewis's formulation in *The Case for Christianity* (though, I don't think this is the strongest formulation of arguments in this vicinity, and I've tried to do a bit better in the main text):

> "Supposing there was no intelligence behind the universe, no creative mind. In that case, nobody designed my brain for the purpose of thinking. It is merely that when the atoms inside my skull happen, for physical or chemical reasons, to arrange themselves in a certain way, this gives me, as a by-product, the sensation I call thought. But, if so, how can I trust my own thinking to be true? It's like upsetting a milk jug and hoping that the way it splashes itself will give you a map of London. But if I can't trust my own thinking, of course I can't trust the arguments leading to atheism, and therefore have no reason to be an atheist, or anything else. Unless I believe in God, I cannot believe in thought: so I can never use thought to disbelieve in God."

[200] See also "Where recursive justification hits bottom" and "My kind of reflection."

the full richness of their values, too.

But just because naturalists *can* do this doesn't mean they *will*. It is, in fact, strangely possible to not be our souls; to cut ourselves off from our full humanity; to live something other than real life. And if we demand that we be only enough soul as we have theory for, I fear we will leave too much of ourselves behind. This isn't to belittle theory, or to sanctify mystery. But just as we can speak before we have a theory of reference; so too can we love past the edges of our theory-of-our-hearts, to the bird with the unknown name; the winter oak; the light on the wall; the splendor of the sky.

Indeed, how much of philosophy is this playing-catch-up, this struggling to understand something you already know, something you were doing-all-along? I love philosophy: but we can't wait to catch up. We never could. But especially not while crossing this bridge, this desert, this new and daunting age. We need to use everything that any part of us knows about goodness and worthiness and holiness and justice. We need to be our souls fully; to carry the thread—even without knowing, fully, what we are carrying.

Herbert writes:

> go because only in this way will you be admitted to the company of cold skulls
> to the company of your ancestors: Gilgamesh Hector Roland
> the defenders of the kingdom without limit and the city of ashes
>
> Be faithful Go

Our city is not ashes yet. The blood still turns in our breasts. We are still chasing with the hounds—and I think the goal is attainable.

But the waters close over us. One of Robinson's characters, a dying pastor, writes a letter to his young son, for the son to read when he grows up farther, and after his father is gone.

> Theologians talk about a prevenient grace that precedes grace itself and allows us to accept it. I think there must also be a prevenient courage that allows us to be brave—that is, to acknowledge that there is more beauty than our eyes can bear, that precious things have been put into our hands and to do nothing to honor them is to do great harm. And therefore, this courage allows us, as the old men said, to make ourselves useful. It allows us to be generous, which is another way of saying exactly the same thing. But that is the pulpit speaking. What have I to leave you but the ruins of old courage, and the lore of old gallantry and hope? Well, as I have said, it is all an ember now, and the good Lord will surely someday breathe it into flame again.
>
> ... I'll pray that you grow up a brave man in a brave country. I'll pray you find a way to be useful.

Chapter X

Loving a world you don't trust

(Warning: spoilers for Angels in America; and moderate spoilers for Harry Potter and the Methods of Rationality.)

> *"I come into the presence of still water..."*
>
> *—Wendell Berry*

A lot of this series has been about problems with *yang*—that is, with the active element in the duality of activity vs. receptivity, doing vs. not-doing, controlling vs. letting go.[201] In particular, I've been interested in the ways that "deep atheism" (that is, a fundamental mistrust towards Nature, and towards bare intelligence) can propel itself towards an ever-more *yang*-y, controlling relationship to Otherness, and to the universe as a whole. I've tried to point at various ways this sort of control-seeking can go wrong in the context of AGI, and to highlight a variety of less-controlling alternatives (e.g. "gentleness," "liberalism/niceness/boundaries," and "green") that I think have a role to play.[202]

This is the final essay in the series. And because I've spent so much time on potential problems with *yang*, and with deep atheism, I want to close with an effort to make sure I've given both of them their due, and been clear about my overall take. To this end, the first part of the essay praises certain types of *yang* directly, in an effort to avoid over-correction towards *yin*. The second part praises something quite nearby to deep atheism that I care about a lot—something I call "humanism." And the third part tries to clarify the depth of atheism I ultimately endorse. In particular, I distinguish between *trust* in the Real, and various other attitudes towards it—attitudes like love, reverence, loyalty, and forgiveness. And I talk about ways these latter attitudes can still look the world's horrors in the eye.

[201] More on this duality here.

[202] Though I've also tried to defend the need for and permissibility of certain types of *yang*—including re: intentionally steering the values of the future. See here and here.

1 In praise of *yang*

Let's start with some words in praise of *yang*.

1.1 In praise of black

Recall "black," from my essay on green. Black, on my construal of the colors, is the color for power, effectiveness, instrumental rationality—and hence, perhaps, the color most paradigmatically associated with *yang*. And insofar as I was especially interested in green qua *yin*, black was green's most salient antagonist.

So I want to be clear: I think black is great.[203] Or at least, some aspects of it. Not black qua ego. Not black that wants power and domination for its sake.[204] Rather: black as the color of *not fucking around*. Of cutting through the bullshit; rejecting what Lewis calls "soft soap"; refusing to pretend things are prettier, or easier, or more comfortable; holding fast to the core thing. I wrote, in my essay on sincerity, about the idea of "seriousness." Black, I think, is the most paradigmatically serious color.

And it's the color of what Yudkowsky calls "the void"—that nameless, final virtue of rationality; the one that carries your movement past your map, past the performance of effort, and into contact with the true goal.[205] Yudkowsky cites Miyamoto Musashi:

> The primary thing when you take a sword in your hands is your intention to cut the enemy, whatever the means... If you think only of hitting, springing, striking or touching the enemy, you will not be able actually to cut him. More than anything, you must be thinking of carrying your movement through to cutting him.

Musashi (image source here)

[203] Though I think that the sort of black I like departs especially much from its connotations in the actual Magic the Gathering universe.

[204] And especially not: black qua more conventional vices like cruelty, contempt, greed, selfishness. Or black qua demons and zombies and corruption and decay (I think that actual Magic the Gathering "black" has a lot of this).

[205] "Every step of your reasoning must cut through to the correct answer in the same movement. More than anything, you must think of carrying your map through to reflecting the territory... If you fail to achieve a correct answer, it is futile to protest that you acted with propriety."

In this sense, I think, black is the color of *actually caring*. That is: one becomes serious, centrally, when there are *stakes;* when one has what Yudkowsky calls "something to protect." And the void is the virtue that won't forget, or half-ass, or look away; that fuels its life with its real fire, and so channels living energy and heat. Indeed, professions of care that seem lacking in black can easily seem like they are missing a mood. Thus, for example, the core push of effective altruism. "Wait, you said that you cared about helping others. So where is the black?"

And because black actually cares, it has *standards*. Green often wants to blur distinctions; to resist binaries and oppositions; to soften and unify and include. But black refuses to unsee the difference between success and failure, excellence and incompetence, truth and falsehood. In this, it aspires to a kind of discipline; some archetypally military virtue; the sort of vibe that emerges when, if you're wrong, then your friends all die.

At various points in this series, I've worried about losing touch with this vibe. It's easy, when writing about green, and Otherness, and tolerance, and "not seeking control," to say soft and blurry and pretty things; kumbaya things. It's easy to play for a certain kind of sage nod; easy to channel parts of the zeitgeist suspicious of archetypically masculine vices in particular, and to get lazy and rose-colored about the full picture. I hoped that the image of a grizzly bear eating Timothy Treadwell alive could help, here. I wanted to remember, up front, about teeth, and blood, and the costs of the wrong *yin*.

And having written so much about *yin*, green, etc, I find myself wanting to praise other black-like virtues as well: virtues like *strength*, health, energy, abundance.[206] I find myself wanting to talk about the enormous benefits of growth and technology; about the ways wealth and power need not be zero-sum; about how much good *yang* the future could have. You've heard this stuff before. I won't dwell. And green can like these things too. But we should remember how much of what we like here is black.

1.2 Bad things are bad

But beyond black qua seriousness, void, discipline, strength, there's something else yang-like I want to honor, here too—something that I associate, more directly, with deep atheism in particular.

Earlier in the series, I quoted Yudkowsky's essay about his brother Yehuda's death.

> ... Yehuda did not "pass on". Yehuda is not "resting in peace". Yehuda is not coming back. Yehuda doesn't exist any more. Yehuda was absolutely annihilated at the age of nineteen. Yes, that makes me angry. I can't put into words how angry. It would be rage to rend the gates of Heaven and burn down God on Its throne, if any God existed. But there is no God, so my anger burns to tear apart the way-things-are, remake the pattern of a world that permits this."

[206] Aaron Gertler informs me that in actual Magic the Gathering, these things are more associated with green than with black. And fair enough. But we should be clear about how closely the *power* that these things grant is tied up with their appeal.

I think the essay about Yehuda may be the best thing Yudkowsky's ever written. If the void is the virtue that channels "real fire," we see that fire burning here. And insofar as the void, in Yudkowsky's portrayal, tries always to "cut the enemy," we see that enemy, too. Or at least, one enemy. "One point eight lives per second, fifty-five million lives per year.... Yehuda's death is the first time I ever lost someone close enough for it to hurt. So now I've seen the face of the enemy. Now I understand, a little better, the price of half a second."

Green is often shy about that word, "enemy." It is suspicious of anger; suspicious of the impulse to kill, to end, to banish. The green-like point, often, is not to *defeat* darkness, but rather: to know it; to find yourself in its gaze; to bring darkness and light together; and so to become more whole.[207] Indeed, green often emphasizes the importance of honoring and making space for the archetypally "bad" side of some duality. Death, decay, suffering, loss—for everything there is a season.[208] And no wonder: unlike deep atheism, green often tries to find some trust or holiness or sacredness in Nature—and Nature is full of such darkness.

But deep atheism has no such allegiance to Nature. And so it is free to recognize that bad things can be just bad: they do not also need to be somehow good, or sacred, or profound. This isn't to say that you shouldn't try to understand them—to see how and why they might be tied up with your own heart, and what it loves. And badness rarely comes pure; rarely "just" bad. But even if something is merely bad *overall*: still, sometimes, once you have looked something dark in the eye, and learned its True Name, then the right choice is, in fact, to fight it; to defeat it; and sometimes, if you don't have better options, to kill it.[209] Cancer cells and invading Nazi soldiers are canonical examples here; and see also this description of smallpox (again using "enemy" rhetoric).[210] Indeed: we rarely have the luxury of understanding some darkness fully before we need to decide whether to fight, or kill. But doing so can be the right choice anyway.

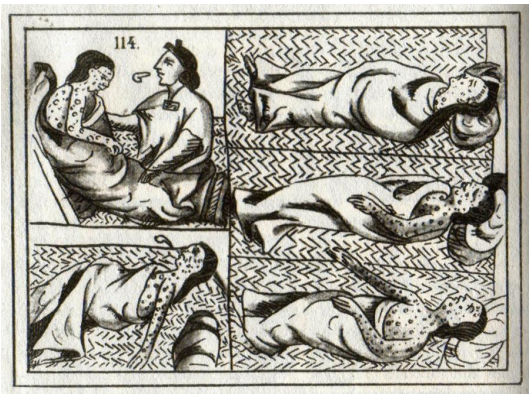
[207] More here.

[208] See e.g. LeGuin's epigraph in the Wizard of Earthsea:

> "Only in silence the word,
> only in dark the light,
> only in dying life:
> bright the hawk's flight
> on the empty sky."

[209] See my discussion in To Light a Candle.

[210] Indeed, any life within nature involves destruction - not just of bacteria and plants and animals and the rest, but also patterns, relationships, possibilities.



16th century Aztec drawing of smallpox victims (Image source here)

Maybe this sounds obvious? "Bad things are bad." But sometimes it breaks into my mind with fresh intensity. I've written, previously, about the way sometimes, on the comparatively rare occasions when I experience even-somewhat-intense suffering or sickness, it comes as a kind of de-fogging. The pain that surrounds me all the time comes rushing back to memory, sharp and sudden—the pain of the people I pass on the street; the pain of the friends who've killed themselves; the pain of hospitals and factory farms, aging and war, dementia and depression and despair. Had I forgotten? But it still seems newly clear, and so central to the story; so core to what's really going on.

At some point, writing this series, I had a mild moment of this flavor. I forget what triggered it. But I lost my taste, in that moment, for something I had been writing about *yin*. I had touched some darkness that reminded me what an indifferent universe means, and I wanted no romance, or cleverness, or evasion. We are more than enough weak and vulnerable. We don't need more. Rather, we need what we're always wanting: more warmth, more light, more strength.

And I had another, related moment visiting Washington D.C., at the World War II memorial. There's a wall fixed with four thousand golden stars, each representing a hundred American soldiers who died in the war. The inscription reads: "Here we mark the price of freedom."

World War II memorial

It's not a *yin* thing. And I don't want to forget.

1.3 Killing dementors

One other note in direct praise of *yang*. There's a scene in Yudkowsky's Harry Potter fanfiction where Harry destroys a dementor. In a sense, I think, it's an extension of the Yehuda essay. And in my opinion, it's the best scene in the book.

Dementor (image source here)

The dementors, Harry has been told, can't be destroyed. The only protection from them is to conjure a ghostly animal—a "patronus"—fueled by thinking happy thoughts. But Harry, initially, can't do the charm. Faced with a dementor, some cold and unspeakable horror crashes through his mind, feeding on him; and the light and goodness inside him almost dies.

But with a friend's help, he survives. And nursed by Dumbledore's phoenix, he looks again, directly, at the horror beneath the tattered cloak—"the void, the emptiness, the hole in the universe, the absence of color and space, the open drain through which warmth poured out of the world." And he sees what he did wrong. Somehow, Harry realizes, the dementors are death—or at least, the shadow of death, cast by magic into the world. "I know you now," he thinks.

> Harry thought of the stars, the image that had almost held off the Dementor even without a Patronus. Only this time, Harry added the missing ingredient, he'd never truly seen it but he'd seen the pictures and the video. The Earth, blazing blue and white with reflected sunlight as it hung in space, amid the black void and the brilliant points of light...
>
> Would they still be plagued by Dementors, the children's children's children, the distant descendants of humankind as they strode from star to star? No. Of course not. The Dementors were only little nuisances, paling into nothingness in the light of that promise; not unkillable, not invincible, not even close. You had to put up with little nuisances, if you were one of the lucky and unlucky few to be born on Earth; on Ancient Earth, as it would be remembered someday. That too was part of what it meant to be alive, if you were one of the tiny handful of sentient beings born into the beginning of all things, before intelligent life had come fully into

> its power. That the much vaster future depended on what you did here, now, in the earliest days of dawn, when there was still so much darkness to be fought, and temporary nuisances like Dementors.

And with this image and others in mind, Harry prepares to think a new kind of happy thought. The patronus charm normally works via the caster blocking out the dementor and thinking about something else; patronuses are animals because their ignorance shelters them from fear. But Harry has trained himself not to take shelter in ignorance, or to look away from darkness. So he looks straight at it instead. He thinks of his utter defiance towards death; of humanity's capacity to end it; and of the way future humans will weep to learn that it ever existed.

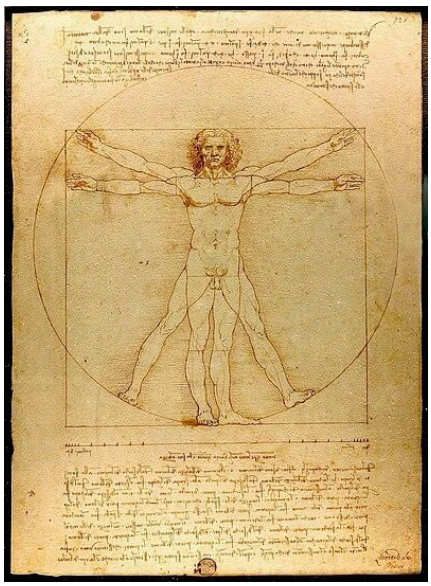
> The wand rose up and leveled straight at the Dementor.
>
> *"EXPECTO PATRONUM!"*
>
> The thought exploded from him like a breaking dam, surged down his arm into his wand, burst from it as blazing white light. Light that became corporeal, took on shape and substance.
>
> A figure with two arms, two legs, and a head, standing upright; the animal *Homo sapiens,* the shape of a human being.
>
> Glowing brighter and brighter as Harry poured all his strength into his spell...
>
> *You are not invincible, and someday the human species will end you.*
>
> *I will end you if I can, by the power of mind and magic and science.*
>
> *I won't cower in fear of Death, not while I have a chance of winning.*
>
> *I won't let Death touch me, I won't let Death touch the ones I love.*
>
> *And even if you do end me before I end you,*
>
> *Another will take my place, and another,*
>
> *Until the wound in the world is healed at last...*
>
> Harry lowered his wand, and the bright figure of a human faded away...
>
> The tattered cloak lay empty within the cage.

The essay about Yehuda named death as enemy. This scene enacts part of the fight, and of the enemy's defeat. And it draws on many of the *yang*-like energies I want to honor: on the void; on defiance and courage; on being willing to look directly at darkness. It channels the thing in the book represented by Godric Gryffindor, and by the cry of the phoenix. Not just black; more like a fusion of red-white-black. Or in the book: red and gold.

To be clear: I'm not endorsing all of Harry's vibe here.[211] Nor am I trying to argue about the merits or mortality of death. But having written this series, and especially the bits about the role of *yin*, I felt some need to point at this scene with the dementor and say "this; this too."

[211] In particular: his talk of the dementors as "little nuisances" smacks, to me, of too much contempt.

2 Humanism



Vitruvian Man, by Da Vinci (Image source here)

Ok, those were some words in praise of *yang*—offered, in part, in an effort to avoid over-correction towards *yin*. In this section, I want to take a moment to do something similar with deep atheism. In particular: having written so much about ways deep atheism can lead to scary places, I don't want to lose touch with something closely related to deep atheism that I care about a lot. I'll call this thing "humanism."

It's an ethic often claimed by atheists, including deep atheists.[212] And while I'll use the term in my own way, I do mean to point at something similar—except, the version I like most.[213] Indeed, as I'll discuss in section 3, the humanism I like can be understood as a specific *form* of deep atheism—except, with a particular sort of existential orientation. This section tries to evoke that orientation directly; in section 3, I talk more explicitly about its contrasts with other forms deep atheism can take.

"Humanism" isn't quite the right word. In particular: it suggests more of a focus on literal humans than I have in mind—something too close to species-ism.[214] Maybe what I really mean is "the project of the Enlightenment." Both terms come with their own baggage; I'll stick with "humanism" for now. But to be clear: humanists can care a lot about non-humans; and non-humans—even "misaligned" ones—could be humanists in the sense I mean.

[212] Indeed, the section in HPMOR with the dementor scene is called "humanism."

[213] There are various "humanist manifestos" out there, but I find that I don't resonate with them so much. And people sometimes think of humanism as a claim about where we should look for *meaning* or *purpose* (i.e., in something human-y, rather than something cosmic or religious); as an approving attitude about humanity's moral character; or as an endorsement of specific sorts of political arrangements and forms of social improvement. These aren't quite what I have in mind. Which isn't to say I disagree with them.

[214] It also conjures contrast with STEM, which is even more off-the-mark.

2.1 Stars and campfires

To get an initial flavor of the sort of "humanism" I have in mind, consider the image, in the dementor scene, of the earth suspended in the dark, amidst the stars.

In my essay on deep atheism, I mentioned the rationalist "secular solstice" event. "Humanist culture," it says on the website. I've been a few times, and it has indeed informed my conception of humanism. And a big theme of the secular solstice is darkness—"a universe that is often cold and uncaring." The image is of winter; of huddling around a campfire; of a night dark and full of terrors. The solstice celebrates the steps humanity has taken out of this night, into light and science and strength. And it looks, like Harry, to the stars.

Do you like the stars? Many humanists do, myself included. Indeed, Harry experiences them with quasi-religious reverence. And see also, classically, Carl Sagan—a central instance of humanism, for me. Here I think, in particular, of classics like "Pale Blue Dot," and (even better, imo) "The Frontier is Everywhere."[215] But also: this scene from the film version of Sagan's novel "Contact," in which the protagonist, sent through space, glimpses some "celestial event" beyond her ship. "No words," she gasps. "They should've sent a poet."[216]

Link to video

But I know people who don't like the stars. The universe, for them, is too big and bleak and cold. And for all the popularity of space stuff amongst deep atheist types, the more negative take on space seems to me the more natural to the worldview. Earth itself is the real campfire, and space the true winter; the longest night; where God's indifference reigns most raw and alien. I thought the movie "Interstellar" did this well; the desolation of the most habitable planets they could find; endless water; endless ice.[217]

This year there was a total eclipse—and for some people, total eclipses disclose this desolation, too. Light can just leave. It gets cold without the sun. Even Annie Dillard, often ecstatic about Nature, found herself unmoored and un-meaning-ed by the black moon:

> There was no world. We were the world's dead people rotating and orbiting around and around, embedded in the planet's crust, while the earth rolled down...

[215] Plus the "Cosmos" series as a whole.
[216] In this version of the script, she adds: "Oh, Palmer, I wish I'd had a baby."
[217] Or at least, the first two the movie visits.

> It had nothing to do with anything. The sun was too small, and too cold, and too far away, to keep the world alive. The white ring was not enough. It was feeble and worthless...
>
> We had all died in our boots on the hilltops of Yakima, and were alone in eternity. Empty space stoppered our eyes and mouths; we cared for nothing.

Image source here

Is it Lovecraft's most famous line? "We live on a placid island of ignorance in the midst of black seas of infinity..." I'll talk about some of my disagreements with Lovecraft below. But I agree that we live on some island, amidst some black sea. And I think of humanism as, partly, about standing together on this island; nurturing our campfire; learning to see into the dark, and to voyage further.

And again, writing about green, I've worried that something about this black sea would get lost, or downplayed. It's not just that earth's forest is brutal and bloody; it's that the forest is *campfire* compared to that dark void, the true wild. Maybe green trusts in the Universe; but the Universe itself is notably un-green, color-wise.[218] So too "Nature." Green is not God's color—not now. Rather, it's a thin film coating a mote of dust.

2.2 Adulthood

So one key image for me, re: humanism, is this uncaring dark; and of working together to protect some flame.

I also associate humanism with something like *adulthood*. Standing on your own feet; looking out of your own eyes; stepping into the real world, and taking responsibility for what you are doing and why.[219]

[218] At least, so far.

[219] See my essay "Seeing more whole" for more on this vibe; and also here and here.

I mentioned the Enlightenment above. The archetypal intellectual vibe I associate with the Enlightenment involves some sense of waking up, growing up, getting a grip. And also, of excitement; of a world newly fresh and open and to-be-lived-in.[220] "Emergence," as Kant put it, from "self-imposed immaturity."

In this sense, I associate humanism with some notion of "dignity"—some sense of a straighter back, and a steadier gaze. I also associate it with a sense of various mediating stories and abstractions falling away; of being left, more, with the raw thing. And I think of it as related to being "alone"; of having, Godric Gryffindor puts it, "only nothingness above."

This last bit sounds a lot like atheism; but actually, I'm not sure.[221] At the least, various enlightenment thinkers were theists of a kind. And atheists often speak about how, even if there *were* a creator God (is there?), he would come with no intrinsic authority; we would still need to judge him for ourselves. Indeed, various humanist fictions feature a God who ends up warranting defiance—see, e.g. *His Dark Materials* (Pullman: "My books are about killing God"); and also *Angels in America*, discussed below.[222] Some sort of theism is true in those books; but I would still call their heroes humanists. In this sense, humanism in my sense is more about how you "look back" at Reality, rather than about what you see.

"Astronomer Copernicus; or Conversations with God" by Jan Metajko (image source here)

[220] Not saying that the archetypal vibe I'm imagining is true to history.

[221] In what direction, exactly, does Godric find nothingness? Where is this empty "above"? And would God need to live there?

[222] I'm counting Angels in America as humanist and atheist. I'm not entirely sure Kushner would.

2.3 Angels in America

As a final pointer at humanism, I want to talk a bit about Tony Kushner's play *Angels in America*—one of my favorite plays,[223] and another paradigm of humanism for me.[224]

Prior Walter is a gay man living in New York City in the 80s. He has AIDS, at a time when AIDS was a death sentence. We see his lesions. We see his friends dying around him. We see him choking down pills by the fistful; collapsing in his hallway; shitting blood.

But also: Prior is having visions. An angel visits him. God, she tells him, has abandoned heaven. The fabric of the world is starting to unravel, and the apocalypse is coming. The angel declares Prior a prophet of *statis*—an end to humanity's movement, migration, exploration—and of death.

But Prior rejects the angel's mission. When she first arrives, he tries to kick her out of his apartment. Later, he wrestles her to the ground, and demands that she take back her prophetic mission, and bless him. Eventually, he ascends to heaven, where he meets the rest of the angels, who try to convince him that death would be a mercy.

Link to video

> ANGEL: We are failing, failing. The Earth and the Angels ... Who demands: More Life, when Death like a protector blinds our eyes, shielding from tender nerve more horror than can be borne? Let any Being on whom Fortune smiles creep away to Death before that last dreadful daybreak, when all your ravaging returns to you...

But Prior refuses.

> PRIOR: But still. Still. Bless me anyway. I want more life. I can't help myself. I do. I've lived through such terrible times, and there are people who live through much much worse, but ... You see them living anyway. When they're more spirit than body, more sores than skin, when they're burned and in agony, when flies lay eggs in the corners of the eyes of their children, they live... We live past hope. If I can find hope anywhere, that's it, that's the best I can do. It's so much not enough, so inadequate but . . . Bless me anyway. I want more life.

[223] Indeed, one of my favorite pieces of art, period.

[224] I specifically love the HBO Miniseries, with Meryl Streep and Al Pacino. There's also a filmed version of the play on the National Theatre website.

Jacob wrestling with the angel (Image source here)

I think some kind of humanist vibe shines through hard in this scene—and elsewhere in the play as well. I mentioned "defiance" above—including towards God himself. When he first gets to heaven, Prior tells the Angels:

> "God—He isn't coming back. And even if He did ... if He ever did come back, if He ever *dared* to show His face ... If after all this destruction, if after all the terrible days of this terrible century He returned to see ... You should *sue* the bastard."

And as ever, the problem with God is evil—pain, disease, loss. There's a scene with a Mormon woman named Harper, who has learned that her husband, Joe, is gay, and that he doesn't love her. At a Mormon visitor's center, the wife in a diorama of a Mormon family crossing the prairie comes to life. Harper speaks to her.

Link to video

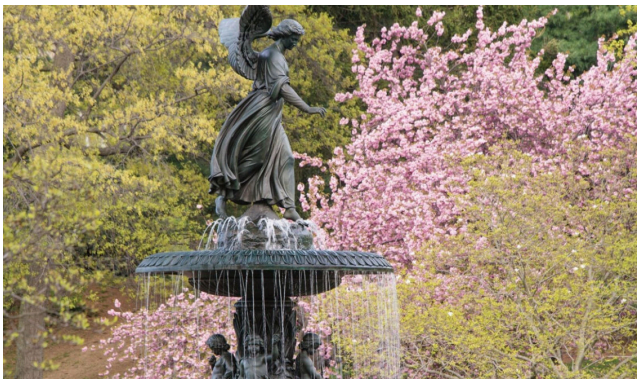
> HARPER: In your experience of the world. How do people change?

> MORMON MOTHER: Well it has something to do with God so it's not very nice. God splits the skin with a jagged thumbnail from throat to belly and then plunges a huge filthy hand in, he grabs hold of your bloody tubes and they slip to evade his grasp but he squeezes hard, he insists, he pulls and pulls till all your innards are yanked out and the pain! We can't even talk about that. And then he stuffs them back, dirty, tangled and torn. It's up to you to do the stitching.

But amidst all this pain, and all this anger at God, the play wants to stand upright, and to find its way. At one point, Harper meets Prior in heaven. She, too, could abandon earth;

but she goes back—devastated by her loss; but fueled, too. "I feel like shit but I've never felt more alive... I don't think God loves His people any better than Joe loved me. The string was cut, and off they went. Ravaged, heartbroken, and free."

And the play believes in what Harper calls "a kind of painful progress." The last scene takes place at a fountain in Central Park. It's an image, we learn, of the biblical fountain of Bethesda—said to flow again at the end of days.[225] "If anyone who was suffering, in the body or the spirit, walked through the waters of the fountain of Bethesda, they would be healed, washed clean of pain."



Bethesda Fountain

The play ends with Prior addressing the audience directly:

Link to video

> PRIOR: This disease will be the end of many of us, but not nearly all, and the dead will be commemorated and will struggle on with the living, and we are not going away. We won't die secret deaths anymore. The world only spins forward. We will be citizens. The time has come.
>
> Bye now. You are fabulous creatures, each and every one. And I bless you: More Life. The Great Work Begins.

[225] I think the reference is to the thing in footnote B of John 5:4 here, though Kushner includes some other backdrop on the angel of Bethesda that I wasn't able to find easily online. The fountain itself was designed by Emma Stebbins, and is purported to be modeled after Charlotte Cushman, Emma's lover and partner, who Emma cared for through her breast cancer.

What is this great work? I don't think it's just beginning. I think it began a long time ago, and we live in its midst. It's the work of campfire, garden, healing water. To unbend our backs. To make gentle the life of this world.[226]

It's real work. We've done it, some. We can do it more.[227] And the future could be, indeed, as fountain.

3 What depth of atheism?

OK, I've now offered some words in praise both of *yang*, and of humanism. In each case, I've done so in an effort to make sure that I don't let core stuff I care about get lost in the series' talk of *yin*, green, and the rest. In both cases, though, I've been channeling something at least nearby to deep atheism fairly hard—despite having pushed back on deep atheism, in various ways, throughout the series. So I want to close with an effort to be clear about the depth of atheism I ultimately endorse.

Deep atheism, as I defined it, was about a fundamental mistrust towards Nature, and towards bare intelligence. Some of that mistrust, I argued, comes from the structure of epistemology itself. Scout mindset accepts that Reality, the ultimate Uncontrolled, could be as arbitrarily horrible as is compatible with your evidence. In that sense, it renounces *a priori* trust—the sort of trust that knows, before looking, that it lives in the arms of a good God, and can rest. Deep atheism admits no such comforts.

This bit seems clearly right to me. But deep atheism, in my discussion, went further. In particular: it drew more specific empirical lessons about which things are what degree trustworthy; it came in with pessimistic priors about whether to expect the Real to be Good; and it endorsed anti-realism about meta-ethics, which made Intelligence orthogonal to Goodness in the same way Nature is—since Intelligence is just Nature, organized and amplified.

My takes on these bits of deep atheism are somewhat more complicated. I agree, obviously, with the empirical basics with respect to death, suffering, the brutality of Nature, and so on. And I do think these are enough to break certain kinds of trust-in-the-Universe. But exactly what types, in which contexts, is a subtler and more detailed question—one I think best approached case-by-case, with "priors" swiftly becoming vastly less relevant. And while meta-ethical anti-realism is by far my best-guess view, I'm less confident in it than some deep atheists I know, and I care about making sure that in worlds where some sort of moral realism is true, we end up in a position to notice this and respond appropriately.[228]

Still, overall, and modulo the messiness of actual empirical forecasting, I'm quite sympa-

[226] Indeed, we've already seen progress on some of the problems that haunt the play. For example: the play is set in 1985, the year the depletion of the ozone was announced—and Harper fixates on the apocalypse it portends. But decades later, we are healing the ozone. And global deaths from HIV/AIDS have halved since 2005. "So much not enough"—yes. But humanist victories nonetheless.

[227] Except, maybe, on cosmic scales.

[228] See e.g. here, here and here for more.

thetic to deep atheism's basic take on the trustworthiness of Nature, and of bare intelligence—where by trustworthiness I mean something like "can be counted on to end up good"; "can be safely taken as an object of *yin*." When I've written, in this series, about "gentleness" and "liberalism/niceness/boundaries" and even about "green," I've meant to be pointing, mostly, at vibes and values that I think are *compatible* with sophisticated (albeit, less paradigmatic) forms of deep atheism, even if more simplistic forms tend to carry momentum in the opposite direction.

But even once you have fixed your degree of *trust* in something, and made your forecasts about how it will behave, this still leaves many other aspects of your overall attitude towards it unresolved. Maybe you do not trust that bear enough to leave your bear mace behind; but does that mean you see its eyes as dead? Maybe you don't trust your five-year-old son to handle your finances; but don't you love him all the same? Maybe you'd die trying to climb that mountain; but is it not beautiful?

So really, deep atheism qua "claim about the universe's trustworthiness" can splinter into a variety of different, more holistic existential orientations. And about these, I'm more opinionated. In particular, I notice that I have use for words like "sacred" and "holy"; for "spirituality"; and for vibes nearby to "green," in ways that I think deep atheists often don't. The essay about "attunement," especially, was trying to point at this bit. Insofar as a given form of deep atheism is on board with that essay—well, then, OK. But insofar as it isn't, or if its paradigmatic vibe isn't, I want to notice the difference.

3.1 The Lovecraft-Sagan spectrum

Can we say more about what this sort of difference consists in? I wrote, in the deep atheism essay, about spirituality as expressing what I called "existential positive."[229] Even without a Big-Man-God, it still turns towards the Real, the Ultimate, with some kind of reverence and affirmation. I think my relationship to the Real has some flavor like this. I don't *trust* the Real to be good. But for all its indifference, it also doesn't land, for me, as neutral, or blank. Rather, the Real has some kind of shine and charge. It calls, wild and silent and too loud to be heard, from some ultimate depth. And the experiences I care about most present themselves as movements in its direction. Hence, indeed, my opposition to experience machines.[230]

To get more of a flavor of what I mean by "existential positive," consider the contrast with H.P. Lovecraft, mentioned above.

[229] See also "Problems of evil" here.

[230] Of course, many atheists-types oppose experience machines as well, even with their altruistic goals secure. But I think that sometimes, at least, this is their spirituality showing through.

(Image source here.)

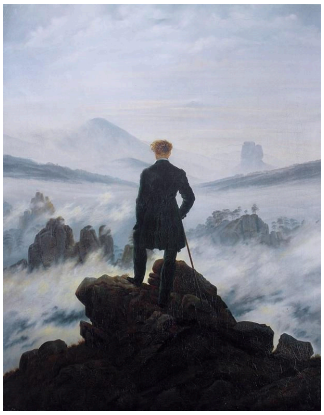
Lovecraft, in my view, is existential *negative*. Here, for example, is the full version of the black sea quote:

> "The most merciful thing in the world, I think, is the inability of the human mind to correlate all its contents. We live on a placid island of ignorance in the midst of black seas of infinity, and it was not meant that we should voyage far. The sciences, each straining in its own direction, have hitherto harmed us little; but some day the piecing together of dissociated knowledge will open up such terrifying vistas of reality, and of our frightful position therein, that we shall either go mad from the revelation or flee from the light into the peace and safety of a new dark age."

Fun stuff. And it's not quite *not* spirituality. At the least, it's not *neutral* on the Real. Rather, the Real fills Lovecraft with a kind of horror. No wonder, then, that he endorses a kind of experience-machining. The night is too dark, and full of terrors; we should swaddle ourselves in some cocoon, and try to forget.

The AI safety scene sometimes draws on Lovecraft's theology (see e.g. shoggoths; or Alexander on Cthulhu). And it's a vibe worth grokking. But I've never gotten into Lovecraft's actual writing, despite multiple attempts.[231] It's not just unpleasant and baroque. Rather, it's like he's looking at a different and less beautiful world. He's too far on the "terror" aspect of sublimity. Everything makes him shudder; everything beckons towards insanity. He'll stand as wanderer over the sea of fog; but instead of splendor he'll see hideous madness.

[231] Specifically, I've now read "The Call of Cthulhu," "The Other Gods," and "At the Mountains of Madness"; recommendations for better stuff welcome.



"The horror, the unspeakable madness ..." (Image source here)

Of course, Lovecraft's is only one form of existential negative. There are many ways to find the Real repugnant; to end up *alienated*, fundamentally, from the world. Still, I think of him as an interesting paradigm. Indeed, I sometimes think of a kind of hazy spectrum between Lovecraft's atheism, and Carl Sagan's. Both sides stand in the enveloping cosmic dark. But the Lovecraft side stands in horror, and the Sagan side, in wonder.[232]

And what about the middle? The middle, as I think of it, stands in a kind of blank. It has no relationship to the Real *per se*. The Real is neutral. Maybe even: boring.

This is the part of the spectrum I associate, mostly directly, with "secularism," and with related forms of "disenchantment." And it's the type I associate with a more watery and domesticated humanism that I *don't* like so much—a type that says something like: "Enough with this cosmic stuff—it's gone dead. But let's enjoy a nice afternoon, and our tea, before it gets cold." Here I think of a talk I heard at an atheist club in undergrad, in which the speaker suggested that in the place of the orienting meaning that religion provides, maybe atheism could promote an activity like ultimate frisbee, which is fun and creates community.

Can you see the difference from Sagan and Kushner and Yudkowsky—and indeed, from Lovecraft? I like tea and frisbee fine. But some kind of existential intensity is getting lost, here. There is some un-relating to the whole story; some blinkering of the attention.

Of course, not all "neutrality" towards the Real need take so tepid a form. And a subtler spectrum, more broadly, would admit many more dimensions, to better capture the many varieties of atheistic passion and indifference.[233] Still, insofar as this simplified spectrum hazily represents different ways of doing deep atheism, I end up on the "positive" end; the Sagan side.

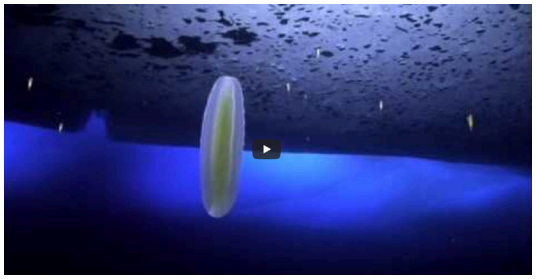
And interestingly, I think that various of the deep atheists I've mentioned in the series

[232] I'm reading Sagan, here, as spiritual in the existential-positive sense; and he thought so, too.

[233] Maybe, for example, we should really be putting "existential-ness" (i.e., the breadth/encompassing-ness of the thing being assigned meaning) and "valence" (e.g., whether the meaning is positive/negative/neutral) on different axes, such that we can better distinguish between a tepid, disenchanted secularism (neutral, but low existential-ness) and a vast and encompassing nihilism (neutral, high existential-ness). Or perhaps we should have three dimensions: existential-ness, valence, and *intensity*.

do, too—at least sometimes. Re: Yudkowsky, for example, I already mentioned Harry's intensely spiritual relationship to the stars; and when Yudkowsky talks about value, he often talks about "minds that can look out at the universe in wonder."[234]

And Herzog, too, for all his pessimism about the brutality of Nature, still holds its beauty in reverence.[235] See, for example, Encounters at the End of the World, in which Herzog travels to Antarctica, partly for the wild desolation, and partly to meet the scientists and workers drawn to the edge of the map. It's an intensely humanistic film, in my view, despite various pessimisms; and it treats the landscape, and the voyaging humans, with a kind of holiness. See, for example, this scene of divers descending under the ice, into what they often call the "cathedral," and the piercing music Herzog places in the background. "To me," says Herzog, "they were like priests preparing for mass."[236]



Link to video

3.2 Mother love and loyalty

But if we are doing "existential positive" in some non-theistic sense—what sense, exactly? And in particular, what does this sense say about "bad things are bad"? If you have reverence towards the Real, does that mean you have reverence towards cancer, rape, dementia, genocide? Aren't you really only "positive" towards particular bits of the Real, rather than the Real itself?

I wrote an essay about this a while back. I do think that the first pass answer is that we are most paradigmatically "positive," even in "spiritual" contexts, towards specific bits of the Real. For all that Ginsberg proclaims everything holy—cock and ass, shit and sand and saxophone—in practice, we direct our reverence via specific centers of meaning and beauty: stars, music, ancient abbeys, light in the trees; and not greasy forks, hemorrhoids, parasites, plaques in your nerve tissue. And even to the extent we are directing our reverence *via* centers of beauty/meaning towards some larger whole, this reverence can't, unconditionally, indicate something like "net good," "better-than-nothing," "I want more of this"—because the larger whole could be not-that-way. Indeed, per scout mindset, we could yet discover that the world is as arbitrarily horrible as is compatible with our evidence—and would we still be "existential positive" then? Are any universes *not* glorious?

[234] See e.g. here and here.

[235] I wrote about Herzog's relationship to nature in the first essay in the series.

[236] The film ends with a line from one of Herzog's interviews: "We are the witness through which the universe becomes conscious of its glory, of its magnificence." It's from this worker, though not in that video.

Bruegel imagines the mouth of hell (Image source here)

We can talk, if we want, about spiritualities that find *some* glory, at least, even in the most un-glorious hell. Or more exotically, about spiritualities that deny that hell is truly *possible*—the way, perhaps, that mathematics *had* to be this beautiful. But I think it's fairly clear that many core aspects of spirituality—especially non-theistic spirituality—have some more restricted and conditional component. They don't, really, revere bare Being (hell has *that*); rather, they revere Being in some particular form.[237]

But importantly, I think they also tend to revere Being in a particular *way*, or a cluster of ways. And I think these ways are often subtler and more interesting than simply calling some stuff good, and some stuff bad. The Lovecraft-Sagan spectrum isn't just "axiology" in disguise—and I want to hold on to the difference.

For example: in the essay on evil, I talked about what Fromm (1956) calls "father love" and "mother love." (To be clear: these are archetypes that actual mothers, fathers, non-gendered parents, non-parents, etc can express to different degrees. Indeed, if we wanted to do more to avoid the gendered connotations, we could just rename them, maybe to something like "assessment love" and "acceptance love.") Fromm's archetypal father assesses the child's merits, and apportions love accordingly—the child has either met a certain standard, or not.[238] But Fromm's archetypal mother does something more complicated. She comes from some more nurturing and loyal place—some place more resilient to the child's faults; a place that does not push the child away, when she sees imperfection, but rather, stays engaged.[239]

And notably, in the context of deep atheism, this doesn't mean that mother love is any less *conscious* of the child's faults; or that it makes worse predictions. Fromm's mother

[237] It's similar to the way "unconditional love" mostly isn't. As Katja Grace puts it, "it is only conditions that separate me from the worms."

[238] This is the paradigmatic stance of population axiology, in judging whether a world is "net negative" or "net positive."

[239] Katja Grace talks about this push away vs. pull-closer distinction in her essay on "Mid-conditional love."

and father can *trust* the child to the same degree—for example, with the finances. But the mother has some *other* sort of *yin* towards the child that the father does not. Some kind of softness, and ongoing attention. Some still-there-for-you. Some not-giving-up.

Chesterton gestures at something similar when he talks about "loyalty."

> My acceptance of the universe is not optimism, it is more like patriotism. It is a matter of primary loyalty. The world is not a lodging-house at Brighton, which we are to leave behind because it is miserable. It is the fortress of our family, with the flag flying on the turret, and the more miserable it is the less we should leave it. The point is not that this world is too sad to love or too glad not to love; the point is that when you do love a thing, its gladness is a reason for loving it, and its sadness a reason for loving it more... What we need is not the cold acceptance of the world as a compromise, but some way in which we can heartily hate and heartily love it. We do not want joy and anger to neutralize each other and produce a surly contentment; we want a fiercer delight and a fiercer discontent.[240]

I think that Prior, in *Angels in America*, has something like this loyalty. As he leaves heaven, his disease—absent in heaven—returns to him: leg pain, constricted lungs, cloudy vision.[241] He gathers his strength, and tells the angels, calmly: "I'm leaving Heaven to you now. I'll take my illness with me, and. And I'll take my death with me, too. The earth's my home, and I want to go home."

Prior hates, fiercely, the world's pain. But he loves his home fiercely, too. And he won't give up.

Should we ever give up? What about: on hell? Well, at the least, we should *destroy* hell.[242] And while you can love hell's *occupants*, and love what it could've been, you can't, really, love what it *is*—not without betraying your true heart. So some kind of father love has its place, too.[243] Most loyalties need limits; and talk of "unconditional love" can get sloppy fast. But I think the distinctive way that mother love and Chestertonian loyalty relate to the world's faults is worth noticing all the same.

3.3 Innocence and forgiveness

And beyond mother love and Chestertonian loyalty, I think "existential positive" can take on other, richly-textured structures as well. In particular, I'm interested in things like "grace," "tragedy," "innocence, and"forgiveness"—all structured, specifically, in relation to the world's faults; but all expressing a type of love and *yin*; something other than alienation, or turning-away.

Jesus, famously, had a thing about some stuff in this vicinity. And I've always found this one of the most compelling parts of Christianity—the way the cross reaches to the bottom of the world's pain and sin, and holds it all. See, also, the prodigal son. And "love your enemies" is related, too.

[240] From "Orthodoxy," chapter 5.
[241] Kushner's description: "Leg pain, constricted lungs, cloudy vision, febrile panic and under that, dreadful weakness."
[242] Sometimes killing is an act of love.
[243] Indeed, a lot of this essay has been in father love's praise.

"The Crucifixion," by Tinoretto (image source here)

But what sort of love? Here's one variant—though not, I think, the one most directly at stake in Christianity.

I wrote, a while ago, about a way in which it's possible to see, underneath the world's evil, a certain kind of innocence. That essay was interested in evolution in particular: the way Nature's brutality emerges, ultimately, from blind, mute, and unselfish patterns—genes—that snag against the swirl of physics, and get carried forward, building pain and violence around them, with no knowledge of what they do. But the point generalizes—at least, conditional on atheism. That is, absent a mind-like creator, the whole thing blooms, ultimately, out of something silent and beyond-moral. Trace the arc of your anger deep enough, and the bottom falls out. Your mind passes through "enemy" to body, neurons, genes, atoms, and on into some more naked and primordial source, where the fingers of blame falter and un-grip.[244]

It's related to the thing Yudkowsky is talking about, when he says: "I do not believe the universe to be evil, a reply which in these days is called atheism." And while "not evil" can seem like cold comfort relative to "good," I find it makes a difference for me. "Nature's silence is its one remark," writes Annie Dillard. When I can hear that silence beneath the noise, some part of me un-clenches, and I find it easier to love the world, for all its pain.[245]

Of course, even if the ultimate ground of things is innocent, the things themselves might not be. Maybe Nature knows nothing of the pain it creates. But humans, Nature's creatures: they know. And sometimes, they make pain anyway. So any talk of "innocence" needs to not lose sight of guilt. Indeed, Jesus came not to declare innocent, but rather to forgive. But even in response to intentional evil, I think that some mix of "innocence," "forgiveness," and "tragedy" can swirl together, against the backdrop of Nature's silence, in a way that makes love easier.

One of my favorite parts of *Angels in America* is about something like this. Roy Cohn, one of the play's villains, is dying in the hospital of AIDS. And he is being haunted by the

[244]Though I think there is an art to doing this in a way that doesn't deny accountability, agency, responsibility, and so on—more discussion here. And also, which treats *yourself*, too, as a part of the world being seen-through.

[245]The eclipse, for me, evoked this kind of silence. And if there is any way to love hell, I expect this silence would be its source.

ghost of Ethel Rosenberg, who he helped send to the electric chair. She tells him:

> I decided to come here so I could see could I forgive you. You who I have hated so terribly I have borne my hatred for you up into the heavens and made a needle-sharp little star in the sky out of it...
>
> I came to forgive but all I can do is take pleasure in your misery...

Eventually, Roy dies—wholly unrepentant.[246] Belize, Roy's night nurse, wants someone to say Kaddish, the Jewish prayer for the dead. But Belize isn't Jewish, so he tries to recruit Louis, who self-describes as an "intensely secular Jew," to help. But Louis, too, hates Roy Cohn:

> LOUIS: Fuck no! For *him*?! No fucking way! ... I can't believe you'd actually pray for—
>
> BELIZE: Louis, I'd even pray for you. He was a terrible person. He died a hard death. So maybe . . . A queen can forgive her vanquished foe. It isn't easy, it doesn't count if it's easy, it's the hardest thing. Forgiveness. Which is maybe where love and justice finally meet. Peace, at least. Isn't that what the Kaddish asks for?

Louis tries to start the Kaddish, but he can't remember it. Then, from the darkness, the ghost of Ethel Rosenberg emerges, speaking the words of the prayer. And they pray together over Roy's body.

Link to video

I love this scene, and the performance above.[247] There is some building pain and anger in their voices, as they pray. They never lose sight of Roy's faults. Indeed, they both end the prayer with "you son of a bitch." But the pain in their voices pierces past Roy,

[246] In particular: Roy tricks Ethel into some kind of compassion for him. He pretends to be speaking, deliriously, to his mother—asking her to sing to him, and saying how scared and sorry he is. Ethel, pitying, sings him a Jewish folk song—Tumbalalaika—as he pretends to die. Then he springs up and shouts "I gotcha!"—then dies for real, as unrepentant as ever.

[247] There's also another version of the full scene here.

to encompass some larger tragedy—the world that created Roy; the world of AIDS and electric chairs and mothers singing to their dying sons. And it holds that world, Roy included, in some fierce and sacralizing gaze.

We also see a kind of forgiveness—albeit, of a different flavor—in HPMOR.[248] Harry has a chance to hurt a man who has killed and tortured victims untold (and who, like Roy, is unrepentant). There would be, Harry thinks, a kind of justice. But Harry thinks of the stars, and of what future generations—the "children's children's children"—would want him to do. And he concludes that to hurt this man any more than needed would be "a not-necessary tragedy happening to one more sentient being ... just be one more element of the sorrows of ancient Earth so long ago." So he lets any sense that he is obligated to be angry "drain away beneath the starlight"; and he casts a spell infused with grace instead.[249]

Of course, in both these cases, the foe in question has been vanquished. Forgiveness is easier once you've won—and it can seem the wrong response to evil triumphant. But even so: I want to notice this energy of including something bad, and even evil, within some broader and ultimately life-affirming frame. "Tragedy." "Grace."

And indeed, how much will we all need the forgiveness of a good future? How much are we all, already, as slaveholders, Nazis, tyrants, in the children's children's eyes? Whatever grace and love we think we warrant, despite our faults, let us extend its rationale unto others as well. And so, too, unto God.

3.4 Creating God

OK, those were some words about the type of deep atheism I ultimately endorse. It's not necessarily that I trust the world to be good in a way that more paradigmatic forms of deep atheism don't. But I think I relate to the world with some kind of "existential positive" all the same—and I've tried to explain how doing so can be compatible with looking "bad things are bad" in the eye, and without reducing spirituality to population ethics. In particular, I've tried to point at the possible role of stuff like mother love, loyalty, innocence, tragedy, and forgiveness. To be clear: I expect that lots of deep atheists are "existential positive" in this sense, at least sometimes; and if you don't like words like "holy" or "sacred," that's OK too. What I care about, here, is some kind of scope and intensity of meaning—and some way this meaning ends up infused with love, and with a kind of *yin*.

[248] The scene I have in mind is here—though, warning, especially spoiler-ish.

[249] Some kinds of green go even further than this. Princess Mononoke, for example, opens with a boar demon—eaten from the inside by writhing, poisonous worms—attacking a village. A village warrior fatally wounds it, and it falls. The villagers fetch the wise woman, who comes out and bows before the demon's body: "Oh nameless god of rage and hate: I bow before you. A mound will be raised and funeral rites performed on this ground where you have fallen. Pass on in peace, and bear us no hatred." The demon, too, is wholly unrepentant: as it dies—melting, toxic, into the earth—it responds to the wise woman: "Disgusting little creatures. Soon all of you will feel my hate, and suffer, as I have suffered." She doesn't turn away.

I'm not sure I'd go quite as green as Miyazaki, here. Should we give active *honor* to the gods of rage and hatred? What about to cancer, or to genocide? If we defeat these gods, what sort of funeral, exactly, should we give them? But I find the scene interesting in its parallels—and contrasts—with the forgiveness at stake in the other, more directly humanist examples in the main text.

But I haven't talked much about the most basic rationale for "existential positive"—namely, *goodness*. Beauty, love, joy—straight up, and full on.[250] Prior, in the last scene of the play, has been living with AIDS for five years; and we see him, in central park, talk about his love for the sunlight in the winter air, and his desire to survive until summer. He's still facing death; and death, often, brings the straight-up-goodness in life into focus.[251] See also here, from a soldier in *All Quiet on the Western Front*, or here (warning: spoilers and violence), from *American Beauty*.

And clearly, some kind of contact with this goodness is core to most kinds of "spirituality." "Holy," "sacred"—they're not quite the same as "good." But they're not too different, either. I quoted from *Gilead*, a book filled with holiness, in the last chapter. "To acknowledge that there is more beauty than our eyes can bear, that precious things have been put into our hands and to do nothing to honor them is to do great harm."

I've written quite a bit about "straight-up-goodness" in the past. See here, for my take on the profound value of a good life today; see here and here, for my take on just how much bigger and better the future could be. This bit is closer to "axiology"—but also, to the possibility of a more wholehearted "yes" to the Real: of father and mother speaking in unison. After all: grace, forgiveness—OK. But what both parents really want for their children is joy.

"So often I have seen the dawn come and the light flood over the land and everything turn radiant at once..." (quote from Gilead; image source here)

I'm not emphasizing the "straight-up-goodness" case for existential positive, though, because I don't think we yet know how much straight-up-goodness the world holds. We know that there is beauty and pain, both. But we don't know the balance, or the pattern. We've read, only, a tiny part of the story; seen, only, in a mirror dimly. And in this sense, we don't yet know who God really is.

[250] Maybe not *all possible worlds* warrant reverence—but what about this one in particular?
[251] More here.

A lot of this is simple ignorance. In the vast realm of the Real, that black sea, where does the arc of the moral universe truly bend? How deep is that schism between *Is* and *Ought*, *Real* and *Good*? "Eluais the god of flowers and free love and he is terrifying"—but how Elder is he, really, and how strong? We can theorize, but we don't yet know. And while our forests and our history offer clues, they are only fragments of the childhood of one mote of dust.

But also: we are not, merely, as onlookers, or as scientists, in the face of God. Rather, at least some small part of God's nature is up to us. We are creating God as we go. "Children of the universe," yes; but parents, too. Or rather: neither children nor parents, but parts, pieces, aspects: some strange mix of separate and the same. And our choices reverberate, and implicate, in ways we don't always track.[252]

Deep atheism, for all its naturalism, sometimes misses this part. It talks as though we stand apart from God—in judgment, and perhaps, in opposition. God, over there, *yanging* at us; and us, over here, *yanging* back. And in one sense: yes. But in another: the whole thing is God, us included. And just as we don't yet know who God is; so, too, we don't yet know who we are, either—what sort of challenges we will rise to; what sort of light and strength we will find within ourselves; what sort of care we will show towards each other, and towards other Others.

But who we are is not merely "discovered." It is chosen. If we wish to learn that we were good, then let us choose goodness. If we wish to learn that on this mote of dust, at least, the arc of the universe bends towards justice, gentleness, peace; then let us create justice; let us be gentle; let us make peace. The arc's direction, after all, is not a static fact—not now, not from the inside. Rather, the arc of the universe is alive. We are looking out of its eyes, moving its hands, hearing its voice in our ears. And when we choose: our choices, all along, will have been God's nature flowing through. If we wish to find more goodness in God's nature, then, let us choose well.

[252] Including, in my view, some ways that standard conceptions of causation don't capture.

4 Final thoughts

I opened the series with Lincoln's second inaugural: "With malice towards none; with charity towards all; with firmness in the right, as God gives us to see the right..." I chose the quote partly because of the concreteness that questions about otherness and control can take on, in the midst of the sort of war Lincoln was fighting, and the sort of peace he was trying to prepare for. But I also like the way it mixes both *yin* and *yang;* gentle and firm; humility and strength.

Lincoln memorial (Image source here)

Reinhold Niebuhr, who took Lincoln as a model of mature spirituality, often inhabits a similar dialectic, in the context of the Cold War: hawk and dove; "in the battle" and "above the battle"; on our own side, and inhabiting some broader and more inclusive perspective.[253] And the dialectical quality can be frustrating. "On the one hand, this; on the other hand, that"—yes, yes, but *what to do*. "All the colors of the wind"—yes, yes, but *which color here?*[254]

Obviously, we need *yin* and *yang*, both—not in some abstract "balance," but in some particular proportion and shape, attuned to the specifics of the case at hand. I've tried, in this series, to offer a few takes on a few semi-specifics re: the age of AGI.[255] But I freely admit that I've left many of the most difficult questions un-addressed, and that much of my focus has been on sharpening our attunement to the structure and momentum of the discourse as a whole, and to the range of orientations available. I hope that these efforts can be useful to people trying to see the age of AGI more whole. But otherness, control—these are old questions. We've asked them before. We'll ask them again. And the real

[253] See e.g. Erwin (2013) for more on Niebuhr's relationship to Lincoln. And see The Irony of American History for a flavor of the dialectic I have in mind.

[254] Obama, also a Niebuhr fan, sometimes uses rhetoric like this.

[255] For example, re: the virtues of liberalism/niceness/boundaries, and re: the (complex) ethics of influencing the values of the future (here and here).

work of wisdom lies in the case by case.

Still, I think it's important to ask the questions fresh. They're old, yes—but if, indeed, ours is the age of AGI, then much of our age will be dauntingly new. We've bred dogs before, but never built a new species smarter than us. We've taught our children values before, but never gradient-descented values into alien, maybe-sentient minds. We've automated before, but never gone obsolete. We've developed new science and technology before; but never at anything close to the sort of pace superintelligence could make possible. And while we've had contests for power before—never with the power of Lewisian "conditioners" so plausibly at stake.

What's more, we will be doing all of this with the specter of war, violence, tyranny, hovering omnipresent in the background. AI alignment risk, after all, is a story of war, and of tyranny. Indeed, the underlying narrative is of Nature, Mind, God everywhere at war with itself. Agency awakes, looks around—and as soon as it's strong enough, it decides to kill all the Others for the sake of its own power, and to install itself on some final throne. And alignment risk aside, I expect the possibility of human war, and of human tyranny, to be lost on very few.

Some wars are worth fighting. And some violence comes unbidden—from bears, Nazis, paperclippers. But I have indeed been trying, throughout the series, to sow seeds of peace. To conceive of the right liberty, that it may endure whatever tests the age of AGI will bring. To remember about gentleness, and pluralism, and cooperation—about *Elua*'s power and virtue both. And more, to remember something deeper about the full richness and force of our values; "the concrete reality in which to participate is to be truly human"; the thing we see when we straighten our backs, and look out of our own eyes—at the real world, and at each other.

Precious things have been placed into our hands. Garden, campfire, healing water. The great work continues—unfinished, alive. May we do our part.