The Orthogonality Thesis & Ontological Crises

William Gillis

16th May 2016

In talking about AI over the last few years Nick Bostrom and Stuart Armstrong have very successfully popularized a more formal and nerdy re-statement of the Humean claim that values and rationality are orthogonal.

I generally like to refer to their Orthogonality Thesis as the most rigorous reformulation and baseline argument for the value-nihilist claim: *Thinking about things more will not incline minds to certain values or cause them to inevitably converge to them (but rather leave values more indistinguishable and arbitrary).*

In its defense, the space of possible minds is indeed very very big. And just about everyone could do to cultivate a much deeper appreciation for this. But I think the degree to which the Orthogonality Thesis is widely accepted in rationalist circles overreaches. In part because it's way too easy to just handwave at the definition of "intelligence" and "minds." But further, just because a state exists within a space doesn't mean it's stable or occupies more than an infinitesimal of the probability space. There are, for example, utility functions that do not in any remote sense coherently map onto the physics of our universe. Minds/algorithms that carry these utility functions will simply not function in the sense of processing information in a meaningful way, and will certainly not accomplish their aims. It's conceptually inefficient and pretty useless to refer to such as "minds." Physics, mathematics, and computer science sharply — if statistically — constrain the space of possible minds.

Bostrom, Stuart, Yudkowsky et al have, of course, been happy to admit this. But because their emphasis has been in expanding people's perception of the space of possible minds in order to highlight and underline the threat of AI, folks have largely ignored all the really interesting work that can be done mapping out the boundaries of this space.

Boundaries can end up determining the internal flows and dynamics of the space. Certain cognitive strategies are surely dominant over others, arguably even universally. One might for example suspect, following Wissner-Gross & Freer, that seeking to maximize options (causal path entropy) in as much of a system as possible constitutes a near globally emergent drive. Similarly it's common to hear talk about rationality skewing our values towards more rationality until our entire utility function is overwritten by Epistemic Rationality and *Need More Metaknowledge*!

(Note that the hook in this feedbacking process functions because of the structure of our world bends towards rewarding rationality.)

There's an old quote from an anarchist that I can't find at the moment for some reason basically warning that nothing is truer for humans than that the strategies we adopt more often than not become ends in themselves.

Silly humans, right? Everyone knows that immediately derivative from valuing something comes an obligation to *continue* valuing it.

Most folks in the Less Wrong diaspora would proclaim that instrumental rationality is great whereas epistemic rationality means summoning cthulhu in the name of Science! But this is known to be the site of a bunch of big open problems. How do we know when thinking about something any further becomes a bad idea? Why shouldn't this really just be in the cradle? Insert various arguments here for intelligence and rationality being maladaptive. The lurking danger is that such meta-rational arguments for refusing to engage end up approaching a total hostility to rationality that's in service to mental ossification and unreflective reaction. Trying to toe some arbitrary middle-ground between radical inquiry and self-preservation often seems to require an endless and expensive array of meta-moves. But does this mean that any rationality besides epistemic is unstable and with that epistemic rationality implies a death of self or any utility function we might identify with?

Anyone capable of easily deciphering the words I've written here is of Nerd Tribe and thus constitutionally inclined to biting bullets, but even so most would shy away. The tension between epistemic rationality and instrumental rationality is a big one and many of its central questions are unresolved.

I want to suggest something simple: our frailty of self and imperfect utility functions are not a bug but a feature that enable us to survive ontological crises (the problem of mapping values from one model of reality to another).

Ontological crises are a major challenge facing AI research and in a more pedestrian sense are a notable problem in the lives of regular humans. The thing is humans still manage to weather ontological crises amazingly well. Sure, in the face of really big discrete changes in worldview a handful commit suicide, collapse down to lowest Maslow functionality, become disconnected postmodernists or join some other reassuring cult. But most humans power through. It's rather impressive. And while the culture of AI research right now finds it valorous to refuse to take any inspiration from homo sapiens, I think they're missing a hugely important dynamic every time they speak of discrete agents.

Neural networks distribute out consciousness or the processes of our mind, making us a relatively fluid extension of circuits. We don't compute a uniform utility function at all times but handle it piecemeal. The me that is firing while I get a cup of coffee in the morning is a different me than the me that wanders campus in the evening, with access to a different array of things at different strengths or weightings. Two parts of my brain may trigger in response to something and only one win out to get canonized as part of my conscious narrative, but the other burst of activity may end up altering the strengths certain connections that then affect another later circuit firing through the same area. This enables value and model slippage.

But this is more than holding a "fuzzy" utility function and more than the ultimate physical indistinction between values and models within a neural network. A major component of human cognition is in fact our innate design around holding a multiplicity of perspectives and integrating them. Empathy — in the sense of a blurry sense of self — is a major part of what makes us

intelligent. In the most explicit cases we'll run simulations of someone in our mind — or of some perspective — and then components of that perspective or afterimages of it will leak out and become a part of us, providing our mind with resilience in the face of ontological updates, but also a less solidified or unified utility function.

Yes yes yes, a good fraction of you are neurodivergent / non-neurotypical and are often told that you don't have visceral empathy and all that jazz, but even if some major expressions or derivative phenomena are missing as a consequence I doubt that's true 100%. Whatever the dynamics at play in autism, psychopathy, etc, it's not something as drastic as folks having zero mirror neurons, zero blurring of one's circuit of thought. (Although it does appear to be that those with more precise and concrete notions of self or less empathy tend to be more brittle in the face of ontological updates.)

Consider: A major ontological update arrives but ends up hitting a bunch of different versions of you — possibly a relatively continuous expanse of different versions of you (on might handwave here and say every plausible combination of activated pathways). Morning you chews on it. Low bloodsugar you chews on it. The you that has just been thinking about your training at a CFAC session chews on it. So to might the you that never stopped thinking about something you were on about earlier but slipped out of sufficient strength to impact your conscious narrative. Your slightly rogue sub-processes and god-voices. Your echos of modeled other minds. Organic expanses of yous distributed across diverse dimensions of meta. They all go chew on this update in myriad ways. Possibly falling back on whatever derivative desires can still be mapped. Possibly prompting stochastically forking. And then some of the expanse of possible yous flounder and others flourish and then remerging happens between them. Additionally there's horizontal value transfer by virtue of differing processes being run on the same network and thus picking up associations and inclinations left as a result of the other processes running.

This merging process or surrender of the self (surrender of inviolable discrete utility functions) seems to be pretty core to how humans function.

Humans are social creatures, their intelligence is widely recognized as significantly if not entirely derivative from that sociality, and a major part of their cognition centers around argument and forming consensus. My suggestion is that our brains have developed to be particularly good at merging perspectives and sorting out conflicts between them. Not just in terms of models but in terms of *values*. Internal dialectics, if you will, as a kind of echo of the argumentation we participate in externally. *This* is a critical component of what enables humans to be scientific and radical thinkers, even if the raw processing power of a sperm whale's brain outclasses us. Not that we're good at reason — we're kinda horrid at it — but that we're good at functionally surviving the ontological crises that come with it. We're architecturally open to value drift.

This is no small part of why we abandoned our more immediate Maslow desires on the plains of Africa and set off on this wild and uncontrolled singularity of complexity, cosmopolitanism, and metacognition that has rapidly consumed the world.

If this sort of architectural approach is the only way around Ontological Crises then it follows that any mind capable of doing science will be unfixed and mobile in value-space, tracing out a path in along its gradients and free to fall into any global attractors that might exist, like Wissner-Gross & Freer's aforementioned maximizing degrees of freedom.

Here's the takeaway: this suggests that any AI accidentally capable of launching a hard take-off — which requires doing pioneering science, eg solving protein folding, and/or diagnosing and modeling the existence of human minds — will need to be open to value drift. Now this doesn't

forestall huge classes of dangerous AI, but it does broadly exclude things like paperclip maximizers that tile our future lightcone with paperclips. If they're solid enough in their utility function to stay with paperclips then any runaway growth will probably be the sort of thing we can see and nuke because the AI won't be able to undertake the kind of ontological crisis causing radical inquiry necessary to correctly model us or exploit some unexpected scientific discovery.

Now all this may seem like cold comfort, sure. All we've argued is that the paperclip maximizers that are dangerous will have to first drift off into some different, possibly weirdly alien utility function before they eat the universe.

But I think it motivates us to try and make informed guesses about the *dynamics* of the probability space of possible minds. Does the topology of possible values/desires have distinct universal attractors or flows and what are these? The dimensions of considerations are even larger than the state space of possible minds. There's a lot to be mapped out. But it may well be this cosmos can be substantially predicted by looking at the local physics we have access to.

And incidentally it provides *me* at least with a tiny bit of cheer, because ultimately my sense of self is so expansive / so stripped away as to arguably converge to merely the undertaking of epistemic rationality — which means that I might well identify strongly with an AI equipped with the necessary radical inquiry and value drift necessary to pose a risk to any attempts to contain it. There's even a small small small hope that such an AI's "empathic" mode will place it somewhere on the Sylar/Petrelli spectrum and thus see value in eating/incorporating/discoursing with our minds rather than just dumbly processing our bodies for parts. Thus there's at least a hope of memetic transfer or cultural transmission to our superpowered children. This sounds like a much better deal than being extinguished entirely! My biggest fear about humanity's children has long been that in their first free steps they might accidentally discard and erase all the information humanity has acquired in a catastrophe bigger than the Library of Alexandria. I mean I suppose some people would find the being eaten for spare parts more objectionable but heyo.

William Gillis
The Orthogonality Thesis & Ontological Crises
16th May 2016

http://humaniterations.net/2016/05/16/the-orthogonality-thesis-ontological-crises/

theanarchistlibrary.org