Best of LessWrong: July 2014

Best of LessWrong: July 2014

Confused as to usefulness of 'consciousness' as a concept

Years ago, before I had come across many of the power tools in statistics, information theory, algorithmics, decision theory, or the Sequences, I was very confused by the concept of intelligence. Like many, I was inclined to reify it as some mysterious, effectively-supernatural force that tilted success at problem-solving in various domains towards the 'intelligent', and which occupied a scale imperfectly captured by measures such as IQ.

Realising that 'intelligence' (as a ranking of agents or as a scale) was a lossy compression of an infinity of statements about the relative success of different agents in various situations was part of dissolving the confusion; the reason that those called 'intelligent' or 'skillful' succeeded more often was that there were underlying processes that had a greater average tendency to output success, and that greater average success caused the application of the labels.

Any agent can be made to lose by an adversarial environment. But for a fixed set of environments, there might be some types of decision processes that do relatively well over that set of environments than other processes, and one can quantify this relative success in any number of ways.

It's almost embarrassing to write that since put that way, it's obvious. But it still seems to me that intelligence is reified (for example, look at most discussions about IQ), and the same basic mistake is made in other contexts, e.g. the commonly-held teleological approach to physical and mental diseases or 'conditions', in which the label is treated as if—by some force of supernatural linguistic determinism—it *causes* the condition, rather than the symptoms of the condition, in their presentation, causing the application of the labels. Or how a label like 'human biological sex' is treated as if it is a true binary distinction that carves reality at the joints and exerts magical causal power over the characteristics of humans, when it is really a fuzzy dividing 'line' in the space of possible or actual humans, the validity of which can only be granted by how well it summarises the characteristics.

For the sake of brevity, even when we realise these approximations, we often use them without commenting upon or disclaiming our usage, and in many cases this is sensible. Indeed, in many cases it's not clear what the exact, decompressed form of a concept would be, or it seems obvious that there can in fact be no single, unique rigorous form of the concept, but that the usage of the imprecise term is still reasonably consistent and correlates usefully with some relevant phenomenon (e.g. tendency to successfully solve problems). Hearing that one person has a higher IQ than another might allow one to make more reliable predictions about who will have the higher lifetime income, for example.

However, widespread use of such shorthands has drawbacks. If a term like 'intelligence' is used without concern or without understanding of its core (i.e. tendencies of agents to succeed in varying situations, or 'efficient cross-domain optimization'), then it might be used teleologically; the term is reified (the mental causal graph goes from "optimising algorithm->success->'intelligent'" to "'intelligent'->success").

In this teleological mode, it feels like 'intelligence' is the 'prime mover' in the system, rather than a description applied retroactively to a set of correlations. But knowledge of those correlations makes the term redundant; once we are aware of the correlations, the term 'intelligence' is just a pointer to them, and does not add anything to them. Despite this, it seems to me that some smart people get caught up in obsessing about reified intelligence (or measures like IQ) as if it were a magical key to all else.

Over the past while, I have been leaning more and more towards the conclusion that the term 'consciousness' is used in similarly dubious ways, and today it occurred to me that there is a very strong analogy between the potential failure modes of discussion of 'consciousness' and between the potential failure modes of discussion of 'intelligence'. In fact, I suspect that the perils of 'consciousness' might be far greater than those of 'intelligence'.

~

A few weeks ago, Scott Aaronson posted to his blog a criticism of integrated information theory (IIT). IIT attempts to provide a quantitative measure of the consciousness of a system. (Specifically, a nonnegative real number phi). Scott points out what he sees as failures of the measure phi to meet the desiderata of a definition or measure of consciousness, thereby arguing that IIT fails to capture the notion of consciousness.

What I read and understood of Scott's criticism seemed sound and decisive, but I can't shake a feeling that such arguments about measuring consciousness are missing the broader point that all such measures of consciousness are doomed to failure from the start, in the same way that arguments about specific measures of intelligence are missing a broader point about lossy compression.

Let's say I ask you to make predictions about the outcome of a game of half-court basketball between Alpha and Beta. Your prior knowledge is that Alpha always beats Beta at (individual versions of) every sport except half-court basketball, and that Beta always beats Alpha at half-court basketball. From this fact you assign Alpha a Sports Quotient (SQ) of 100 and Beta an SQ of 10. Since Alpha's SQ is greater than Beta's, you confidently predict that Alpha will beat Beta at half-court.

Of course, that would be wrong, wrong, wrong; the SQ's are encoding (or compressing) the comparative strengths and weaknesses of Alpha and Beta across various sports, and in particular that Alpha always loses to Beta at half-court. (In fact, if other combinations lead to the same SQ's, then *not even that much* information is encoded, since other combinations might lead to the same scores.) So to just look at the SQ's as numbers and use that as your prediction criterion is a knowably inferior strategy to looking at the details of the case in question, i.e. the actual past results of half-court games between the two.

Since measures like this fictional SQ or actual IQ or fuzzy (or even quantitative) notions of consciousness are at best shorthands for specific abilities or behaviours, tabooing the shorthand should never leave you with less information, since a true shorthand, by its very nature, does not add any information.

When I look at something like IIT, which (if Scott's criticism is accurate) assigns a superhuman consciousness score to a system that evaluates a polynomial at some points, my reaction is pretty much, "Well, this kind of flaw is pretty much inevitable in such an overambitious definition."

Six months ago, I wrote:

"...it feels like there's a useful (but possibly quantitative and not qualitative) difference between myself (obviously 'conscious' for any coherent extrapolated meaning of the term) and my computer (obviously not conscious (to any significant extent?))..."

Mark Friedenbach replied recently (so, a few months later):

"Why do you think your computer is not conscious? It probably has more of a conscious experience than, say, a flatworm or sea urchin. (As byrnema notes, conscious does not necessarily imply self-aware here.)"

I feel like if Mark had made that reply soon after my comment, I might have had a hard time formulating why, but that I would have been inclined towards disputing that my computer is conscious. As it is, at this point I am struggling to see that there is any meaningful disagreement here. Would we disagree over what my computer can do? What information it can process? What tasks it is good for, and for which not so much?

What about an animal instead of my computer? Would we feel the same philosophical confusion over any given capability of an average chicken? An average human?

Even if we did disagree (or at least did not agree) over, say, an average human's ability to detect and avoid ultraviolet light without artificial aids and modern knowledge, this lack of agreement would not feel like a messy, confusing philosophical one. It would feel like one tractable to direct experimentation. You know, like, blindfold some experimental subjects, control subjects, and experimenters and see how the experimental subjects react to ultraviolet light versus other light in the control subjects. Just like if we were arguing about whether Alpha or Beta is the better athlete, there would be no mystery left over once we'd agreed about their relative abilities at every athletic activity. At most there would be terminological bickering over which scoring rule over athletic activities we should be using to measure 'athletic ability', but not any disagreement for any fixed measure.

I have been turning it over for a while now, and I am struggling to think of contexts in which consciousness really holds up to attempts to reify it. If asked why it doesn't make sense to politely ask a virus to stop multiplying because it's going to kill its host, a conceivable response might be something like, "Erm, you know it's not conscious, right?" This response might well do the job. But if pressed to cash out this response, what we're really concerned with is the absence of the usual physical-biological processes by which talking at a system might affect its behaviour, so that there is no reason to expect the polite request to increase the chance of the favourable outcome. Sufficient knowledge of physics and biology could make this even more rigorous, and no reference need be made to consciousness.

The only context in which the notion of consciousness seems inextricable from the statement is in ethical statements like, "We shouldn't eat chickens because they're conscious." In such statements, it feels like a particular sense of 'conscious' is being used, one which is *defined* (or at least characterised) as 'the thing that gives moral worth to creatures, such that we shouldn't eat them'. But then it's not clear why we should call this moral criterion 'consciousness'; insomuch as consciousness is about information processing or understanding an environment, it's not obvious what connection this has to moral worth. And insomuch as consciousness is the Magic

Token of Moral Worth, it's not clear what it has to do with information processing.

If we relabelled zxcv=conscious and rewrote, "We shouldn't eat chickens because they're zxcv," then this makes it clearer that the explanation is not entirely satisfactory; what does zxcv have to do with moral worth? Well, what does consciousness have to do with moral worth? Conservation of argumentative work and the usual prohibitions on equivocation apply: You can't introduce a new sense of the word 'conscious' then plug it into a statement like "We shouldn't eat chickens because they're conscious" and dust your hands off as if your argumentative work is done. That work is done only if one's actual values and the definition of consciousness to do with information processing already exactly coincide, and this coincidence is known. But it seems to me like a claim of any such coincidence must stem from confusion rather than actual understanding of one's values; valuing a system commensurate with its ability to process information is a fake utility function.

When intelligence is reified, it becomes a teleological fake explanation; consistently successful people are consistently successful because they are known to be Intelligent, rather than their consistent success causing them to be called intelligent. Similarly consciousness becomes teleological in moral contexts: We shouldn't eat chickens because they are called Conscious, rather than 'these properties of chickens mean we shouldn't eat them, and chickens also qualify as conscious'.

So it is that I have recently been very skeptical of the term 'consciousness' (though grant that it can sometimes be a useful shorthand), and hence my question to you: Have I overlooked any counts in favour of the term 'consciousness'?

Confound it! Correlation is (usually) not causation! But why not?

It is widely understood that statistical correlation between two variables ≠ causation. But despite this admonition, people are routinely overconfident in claiming correlations to support particular causal interpretations and are surprised by the results of randomized experiments, suggesting that they are biased & systematically underestimating the prevalence of confounds/common-causation. I speculate that in realistic causal networks or DAGs, the number of possible correlations grows faster than the number of possible causal relationships. So confounds really are that common, and since people do not think in DAGs, the imbalance also explains overconfidence.

Full article: http://www.gwern.net/Causality

[LINK] Claustrum Stimulation Temporarily Turns Off Consciousness in an otherwise Awake Patient

This paper, or more often the New Scientist's exposition of it is being discussed online and is rather topical here. In a nutshell, stimulating one small but central area of the brain reversibly rendered one epilepsia patient unconscious without disrupting wakefulness. Impressively, this phenomenon has apparently been hypothesized before, just never tested (because it's hard and usually unethical). A quote from the New Scientist article (emphasis mine):

> One electrode was positioned next to the claustrum, an area that had never been stimulated before.
>
> When the team zapped the area with high frequency electrical impulses, the woman lost consciousness. She stopped reading and stared blankly into space, she didn't respond to auditory or visual commands and her breathing slowed. As soon as the stimulation stopped, she immediately regained consciousness with no memory of the event. The same thing happened every time the area was stimulated during two days of experiments (Epilepsy and Behavior, doi.org/tgn).
>
> To confirm that they were affecting the woman's consciousness rather than just her ability to speak or move, the team asked her to repeat the word "house" or snap her fingers before the stimulation began. If the stimulation was disrupting a brain region responsible for movement or language she would have stopped moving or talking almost immediately. Instead, **she gradually spoke more quietly or moved less and less until she drifted into unconsciousness**. Since there was no sign of epileptic brain activity during or after the stimulation, the team is sure that it wasn't a side effect of a seizure.

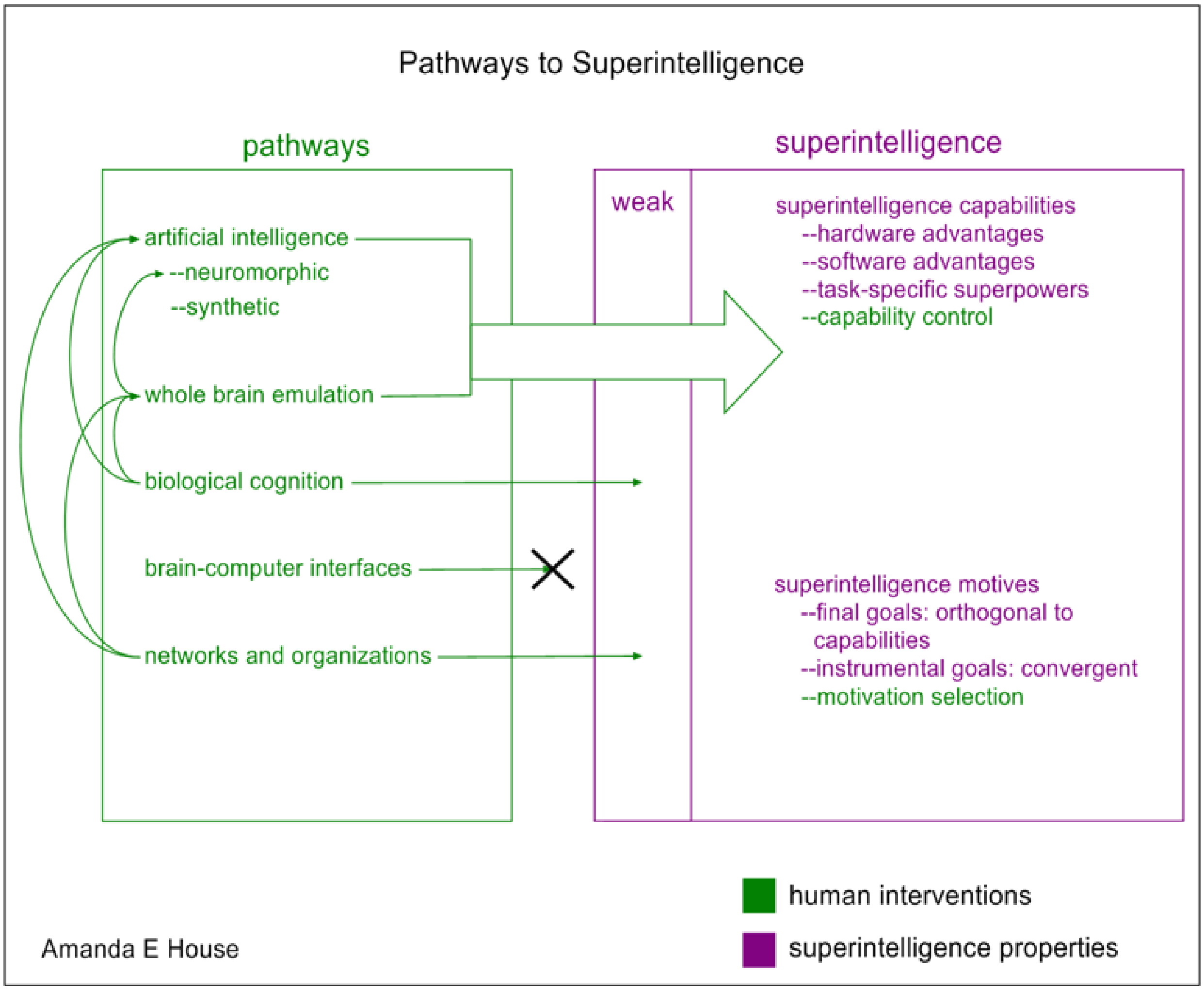
If confirmed, this hints at several interesting points. For example, a complex enough brain is not sufficient for consciousness, a sort-of command and control structure is required, as well, even if relatively small. A low-consciousness state of late-stage dementia sufferers might be due to the damage specifically to the claustrum area, not just the overall brain deterioration. The researchers speculates that stimulating the area in vegetative-state patients might help "push them out of this state". From an AI research perspective, understanding the difference between wakefulness and consciousness might be interesting, too.

A Visualization of Nick Bostrom’s Superintelligence

Through a series of diagrams, this article will walk through key concepts in Nick Bostrom’s *Superintelligence*. The book is full of heavy content, and though well written, its scope and depth can make it difficult to grasp the concepts and mentally hold them together. The motivation behind making these diagrams is not to repeat an explanation of the content, but rather to present the content in such a way that the connections become clear. Thus, this article is best read and used as a supplement to *Superintelligence*.

Note: *Superintelligence* is now available in the UK. The hardcover is coming out in the US on September 3. The Kindle version is already available in the US as well as the UK.

Roadmap: there are two diagrams, both presented with an accompanying description. The two diagrams are combined into one mega-diagram at the end.



Pathways to Superintelligence
pathways
artificial intelligence
--neuromorphic
--synthetic
whole brain emulation
biological cognition
brain-computer interfaces
networks and organizations
superintelligence
weak
superintelligence capabilities
--hardware advantages
--software advantages
--task-specific superpowers
--capability control
superintelligence motives
--final goals: orthogonal to capabilities
--instrumental goals: convergent
--motivation selection
🟩 human interventions
🟪 superintelligence properties
Amanda E House

Figure 1: Pathways to Superintelligence

Figure 1 displays the five pathways toward superintelligence that Bostrom describes in chapter 2 and returns to in chapter 14 of the text. According to Bostrom, brain-computer interfaces are unlikely to yield superintelligence. Biological cognition, i.e., the enhancement of human intelligence, may yield a weak form of superintelligence on its own. Additionally, improvements to biological cognition could feed back into driving the progress of artificial intelligence or whole brain emulation. The arrows from networks and organizations likewise indicate technologies feeding back into AI and whole brain emulation development.

Artificial intelligence and whole brain emulation are two pathways that can lead to fully realized superintelligence. Note that neuromorphic is listed under artificial intelligence, but an arrow connects from whole brain emulation to neuromorphic. In chapter 14, Bostrom suggests that neuromorphic is a potential outcome of incomplete or improper whole brain emulation. Synthetic AI includes all the approaches to AI that are not neuromorphic; other terms that have been used are algorithmic or *de novo* AI.

Figure 1 also includes some properties of superintelligence. In regard to its capabilities, Bostrom discusses software and hardware advantages of a superintelligence in chapter 3, when describing possible forms of superintelligence. In chapter 6, Bostrom discusses the superpowers a superintelligence may have. The term “task-specific superpowers” refers to Table 8, which contains tasks (e.g., strategizing or technology research), and corresponding skill sets (e.g., forecasting, planning, or designing) which a superintelligence may have. Capability control, discussed in chapter 9, is the limitation of a superintelligence’s abilities. It is a response to the problem of preventing undesirable outcomes. As the problem is one for human programmers to analyze and address, capability control appears in green.

In addition to what a superintelligence might do, Bostrom discusses why it would do those things, i.e., what its motives will be. There are two main theses—the orthogonality thesis and the instrumental convergence thesis—both of which are expanded upon in chapter 7. Motivation selection, found in chapter 9, is another method to avoid undesirable outcomes. Motivation selection is the loading of desirable goals and purposes into the superintelligence, which would potentially render capability control unnecessary. As motivation selection is another problem for human programmers, it also appears in green.

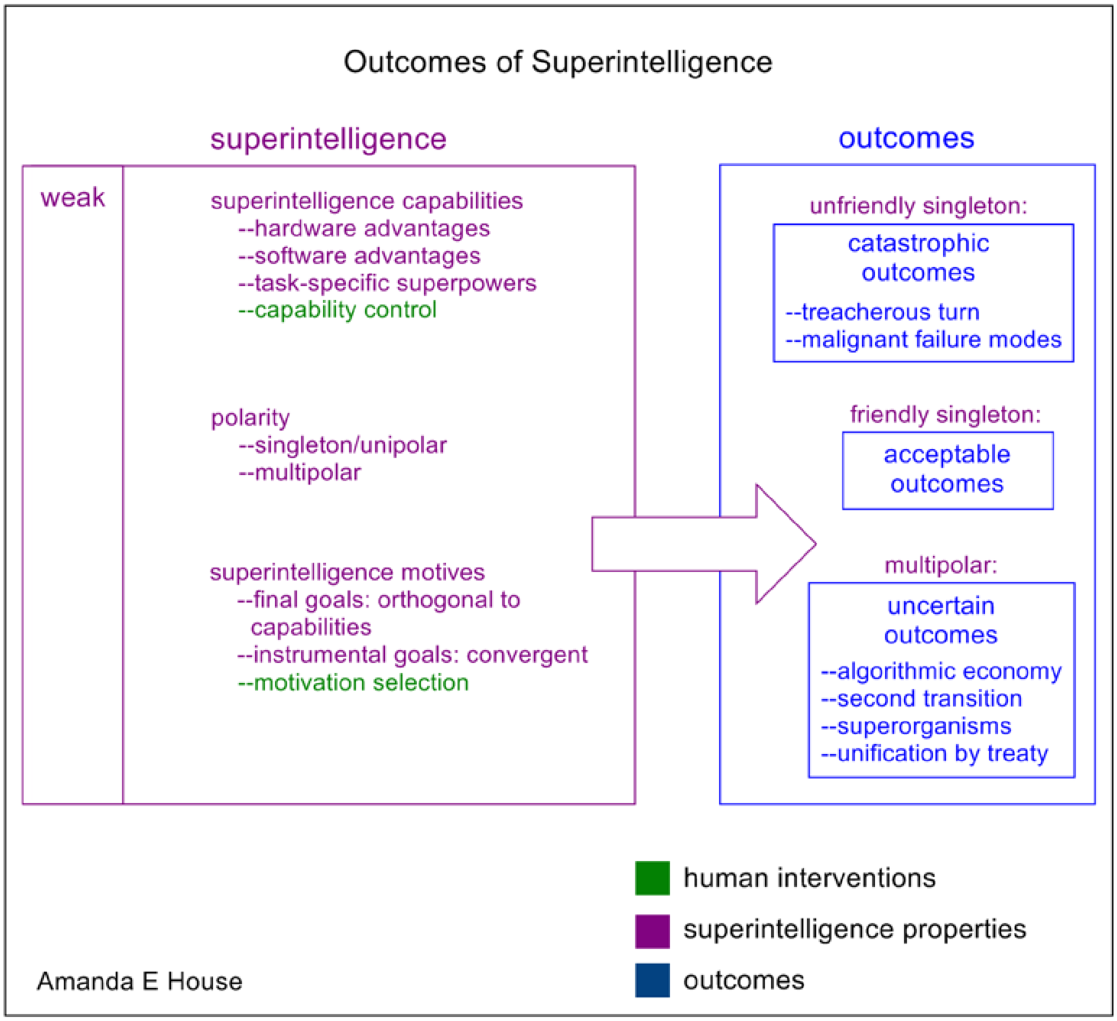


Figure 2: Outcomes of Superintelligence

Figure 2 maps the types of superintelligence to the outcomes. It also introduces some terminology which goes beyond general properties of superintelligence, and breaks up the types of superintelligence as well. There are two axes which divide superintelligence. One is the polarity, i.e., the possibility of a singleton or multipolar scenario. The other is the difference between friendly and unfriendly superintelligence. Polarity is slightly between superintelligence properties and outcomes; it refers to a combination of human actions and design of superintelligence, as well as actions of a superintelligence. Thus, polarity terms appear in both the superintelligence and the outcomes areas of Figure 2. Since safety profiles are a consequence of many components of superintelligence, those terms appear in the outcomes area.

Bostrom describes singletons in the most detail. An unfriendly singleton leads to existential risks, including scenarios which Bostrom describes in chapter 8. In contrast, a friendly superintelligence leads to acceptable outcomes. Acceptable outcomes are not envisioned in as great detail as existential risks; however, chapter 13 discusses how a superintelligence operating under coherent extrapolated volition or one of the morality models would behave. This could be seen as an illustration of what a successful attempt at friendly superintelligence would yield. A multipolar scenario of superintelligence is more difficult to predict; Bostrom puts forth various visions found in chapter 11. The one which receives the most attention is the algorithmic economy, based on Robin Hanson’s ideas.

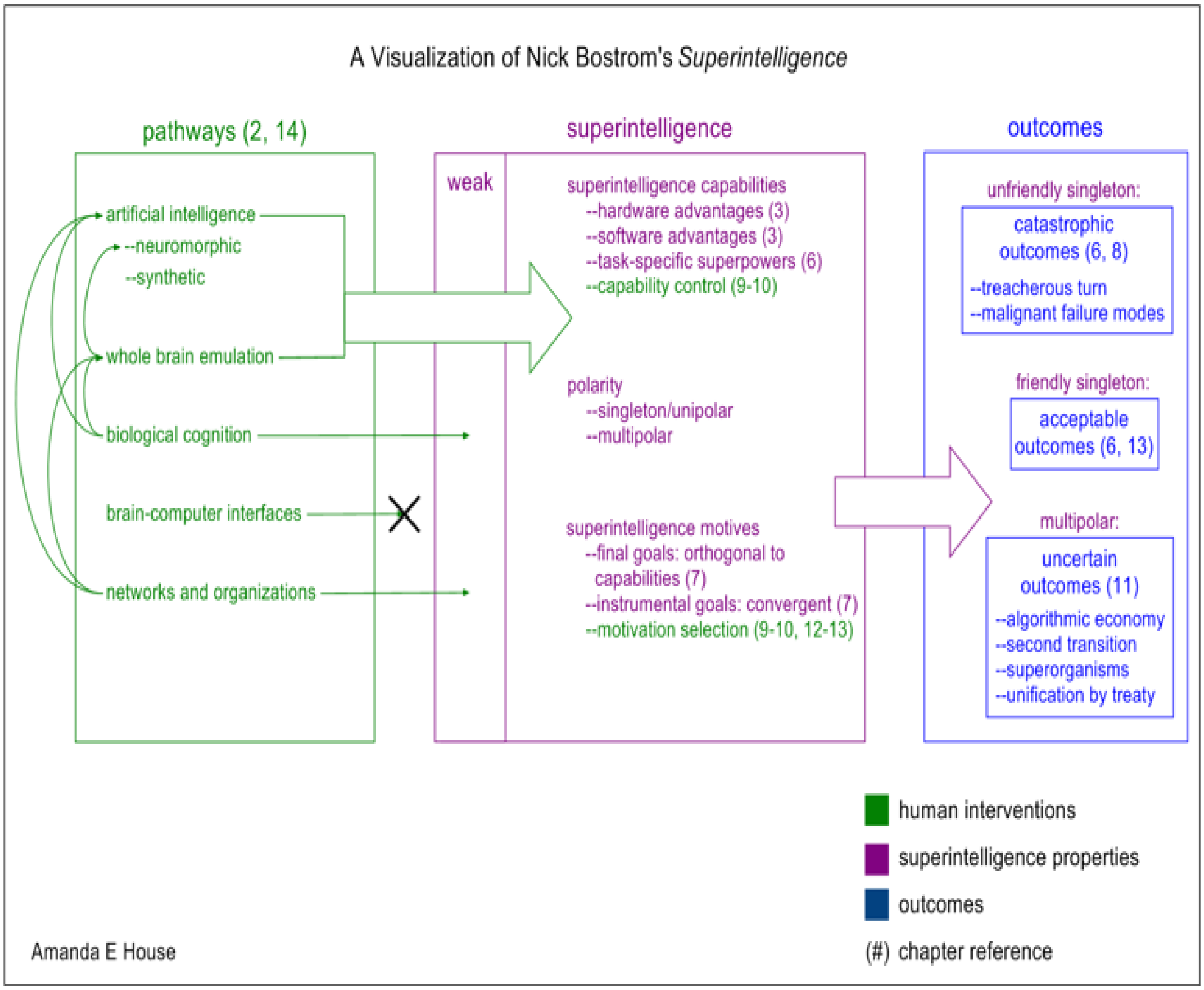

A Visualization of Nick Bostrom's *Superintelligence*
pathways (2, 14)
artificial intelligence
--neuromorphic
--synthetic
whole brain emulation
biological cognition
brain-computer interfaces
networks and organizations
superintelligence
weak
superintelligence capabilities
--hardware advantages (3)
--software advantages (3)
--task-specific superpowers (6)
--capability control (9-10)
polarity
--singleton/unipolar
--multipolar
superintelligence motives
--final goals: orthogonal to capabilities (7)
--instrumental goals: convergent (7)
--motivation selection (9-10, 12-13)
outcomes
unfriendly singleton:
catastrophic outcomes (6, 8)
--treacherous turn
--malignant failure modes
friendly singleton:
acceptable outcomes (6, 13)
multipolar:
uncertain outcomes (11)
--algorithmic economy
--second transition
--superorganisms
--unification by treaty
🟩 human interventions
🟪 superintelligence properties
🟦 outcomes
(#) chapter reference
Amanda E House

Figure 3: A Visualization of Nick Bostrom’s *Superintelligence*

Finally, figures 1 and 2 are put together for the full diagram in Figure 3. As Figure 3 is an overview of the book's contents, it includes chapter numbers for parts of the diagram. This allows Figure 3 to act as a quick reference and guide readers to the right part of *Superintelligence* for more information.

Acknowledgements

Thanks to Steven Greidinger for co-creating the first versions of the diagrams, to Robby Bensinger for his insights on the final revisions, to Alex Vermeer for his suggestions on early drafts, and to Luke Muehlhauser for the time to work on this project during an internship at the Machine Intelligence Research Institute.

Look for the Next Tech Gold Rush?

In early 2000, I registered my personal domain name weidai.com, along with a couple others, because I was worried that the small (sole-proprietor) ISP I was using would go out of business one day and break all the links on the web to the articles and software that I had published on my "home page" under its domain. Several years ago I started getting offers, asking me to sell the domain, and now they're coming in almost every day. A couple of days ago I saw the first six figure offer ($100,000).

In early 2009, someone named Satoshi Nakamoto emailed me personally with an announcement that he had published version 0.1 of Bitcoin. I didn't pay much attention at the time (I was more interested in Less Wrong than Cypherpunks at that point), but then in early 2011 I saw a LW article about Bitcoin, which prompted me to start mining it. I wrote at the time, "thanks to the discussion you started, I bought a Radeon 5870 and started mining myself, since it looks likely that I can at least break even on the cost of the card." That approximately $200 investment (plus maybe another $100 in electricity) is also worth around six figures today.

Clearly, technological advances can sometimes create gold rush-like situations (i.e., first-come-first-serve opportunities to make truly extraordinary returns with minimal effort or qualifications). And it's possible to stumble into them without even trying. Which makes me think, maybe we *should* be trying? I mean, if only I had been *looking* for possible gold rushes, I could have registered a hundred domain names optimized for potential future value, rather than the few that I happened to personally need. Or I could have started mining Bitcoins a couple of years earlier and be a thousand times richer.

I wish I was already an experienced gold rush spotter, so I could explain how best to do it, but as indicated above, I participated in the ones that I did more or less by luck. Perhaps the first step is just to keep one's eyes open, and to keep in mind that tech-related gold rushes do happen from time to time and they are not impossibly difficult to find. What other ideas do people have? Are there other past examples of tech gold rushes besides the two that I mentioned? What might be some promising fields to look for them in the future?

Downvote stalkers: Driving members away from the LessWrong community?

Last month I saw this post: http://lesswrong.com/lw/kbc/meta_the_decline_of_discussion_now_with_charts/ addressing whether the discussion on LessWrong was in decline. As a relatively new user who had only just started to post comments, my reaction was: “I hope that LessWrong isn’t in decline, because the sequences are amazing, and I really like this community. I should try to write a couple articles myself and post them! Maybe I could do an analysis/summary of certain sequences posts, and discuss how they had helped me to change my mind”. I started working on writing an article.

Then I logged into LessWrong and saw that my Karma value was roughly half of what it had been the day before. Previously I hadn’t really cared much about Karma, aside from whatever micro-utilons of happiness it provided to see that the number slowly grew because people generally liked my comments. Or at least, I thought I didn’t really care, until my lizard brain reflexes reacted to what it perceived as an assault on my person.

Had I posted something terrible and unpopular that had been massively downvoted during the several days since my previous login? No, in fact my ‘past 30 days’ Karma was still positive. Rather, it appeared that everything I had ever posted to LessWrong now had a -1 on it instead of a 0. Of course, my loss probably pales in comparison to that of other, more prolific posters who I have seen report this behavior.

So what controversial subject must I have commented on in order to trigger this assault? Well, let’s see, in the past week I had asked if anyone had any opinions of good software engineer interview questions I could ask a candidate. I posted in http://lesswrong.com/lw/kex/happiness_and_children/ that I was happy to not have children, and finally, here in what appears to me to be by far the most promising candidate:http://lesswrong.com/r/discussion/lw/keu/separating_the_roles_of_theory_and_direct/ I replied to a comment about global warming data, stating that I routinely saw headlines about data supporting global warming.

Here is our scenario: A new user is attempting to participate on a message board that values empiricism and rationality, posted that evidence supports that climate change is real. (Wow, really rocking the boat here!) Then, apparently in an effort to ‘win’ this discussion by silencing opposition, someone went and downvoted every comment this user had ever made on the site. Apparently they would like to see LessWrong be a bastion of empiricism and rationality and [i]climate change denial[/i] instead? And the way to achieve this is not to have a fair and rational discussion of the existing empirical data, but rather to simply Karmassassinate anyone who would oppose them?

Here is my hypothesis: The continuing problem of karma downvote stalkers is contributing to the decline of discussion on the site.   I definitely feel much less motivated to try and contribute anything now, and I have been told by multiple other people at LessWrong meetings things such as “I used to post a lot on LessWrong, but then I posted X, and got mass downvoted, so now I only comment on Yvain’s blog”. These anecdotes are, of course, only very weak evidence to support my claim. I wish I could provide more, but I will have to defer to any readers who can supply more.

Perhaps this post will simply trigger more retribution, or maybe it will trigger an outswelling of support, or perhaps just be dismissed by people saying I should’ve posted it to the weekly discussion thread instead.   Whatever the outcome, rather than meekly leaving LessWrong and letting my 'stalker' win, I decided to open a discussion about the issue. Thank you!

This is why we can't have social science

Jason Mitchell is [edit: has been] the John L. Loeb Associate Professor of the Social Sciences at Harvard. He has won the National Academy of Science's Troland Award as well as the Association for Psychological Science's Janet Taylor Spence Award for Transformative Early Career Contribution.

Here, he argues against the principle of replicability of experiments in science. Apparently, it's disrespectful, and presumptively wrong.

> Recent hand-wringing over failed replications in social psychology is largely pointless, because unsuccessful experiments have no meaningful scientific value.
>
> Because experiments can be undermined by a vast number of practical mistakes, the likeliest explanation for any failed replication will always be that the replicator bungled something along the way. Unless direct replications are conducted by flawless experimenters, nothing interesting can be learned from them.
>
> Three standard rejoinders to this critique are considered and rejected. Despite claims to the contrary, failed replications do not provide meaningful information if they closely follow original methodology; they do not necessarily identify effects that may be too small or flimsy to be worth studying; and they cannot contribute to a cumulative understanding of scientific phenomena.
>
> Replication efforts appear to reflect strong prior expectations that published findings are not reliable, and as such, do not constitute scientific output.
>
> The field of social psychology can be improved, but not by the publication of negative findings. Experimenters should be encouraged to restrict their “degrees of freedom,” for example, by specifying designs in advance.
>
> Whether they mean to or not, authors and editors of failed replications are publicly impugning the scientific integrity of their colleagues. Targets of failed replications are justifiably upset, particularly given the inadequate basis for replicators’ extraordinary claims.

This is why we can't have social science. Not because the subject is not amenable to the scientific method -- it obviously is. People *are* conducting controlled experiments and other people are attempting to replicate the results. So far, so good. Rather, the problem is that at least one celebrated authority in the field *hates* that, and would prefer much, much more deference to authority.

Politics is hard mode

Summary: I don't think 'politics is the mind-killer' works well rhetorically. I suggest 'politics is hard mode' instead.

Some people in and catawampus to the LessWrong community have objected to "*politics is the mind-killer*" as a framing (/ slogan / taunt). Miri Mogilevsky explained on Facebook:

> My usual first objection is that it seems odd to single politics out as a “mind-killer” when there’s plenty of evidence that tribalism happens *everywhere*. Recently, there has been a whole kerfuffle within the field of psychology about replication of studies. Of course, some key studies have failed to replicate, leading to accusations of “bullying” and “witch-hunts” and what have you. Some of the people involved have since walked their language back, but it was still a rather concerning demonstration of mind-killing in action. People took “sides,” people became upset at people based on their “sides” rather than their actual opinions or behavior, and so on.
>
> Unless this article refers specifically to electoral politics and Democrats and Republicans and things (not clear from the wording), “politics” is such a frightfully broad category of human experience that writing it off entirely as a mind-killer that cannot be discussed or else all rationality flies out the window effectively prohibits a large number of important issues from being discussed, by the very people who can, in theory, be counted upon to discuss them better than most. Is it “politics” for me to talk about my experience as a woman in gatherings that are predominantly composed of men? Many would say it is. But I’m sure that these groups of men stand to gain from hearing about my experiences, since some of them are concerned that so few women attend their events.
>
> In this article, Eliezer notes, “Politics is an important domain to which we should individually apply our rationality — but it’s a terrible domain in which to learn rationality, or discuss rationality, unless all the discussants are already rational.” But that means that we all have to individually, privately apply rationality to politics without consulting anyone who can help us do this well. After all, there is no such thing as a discussant who is “rational”; there is a reason the website is called “Less Wrong” rather than “Not At All Wrong” or “Always 100% Right.” Assuming that we are all trying to be more rational, there is nobody better to discuss politics with than each other.
>
> The rest of my objection to this meme has little to do with this article, which I think raises lots of great points, and more to do with the response that I’ve seen to it — an eye-rolling, condescending dismissal of politics itself and of anyone who cares about it. Of course, I’m totally fine if a given person isn’t interested in politics and doesn’t want to discuss it, but then they should say, “I’m not interested in this and would rather not discuss it,” or “I don’t think I can be rational in this discussion so I’d rather avoid it,” rather than sneeringly reminding me “You know, politics is the mind-killer,” as though I am an errant child. I’m well-aware of the dangers of politics to good thinking. I am also aware of the benefits

> of good thinking to politics. So I’ve decided to accept the risk and to try to apply good thinking there. [...]
>
> I’m sure there are also people who disagree with the article itself, but I don’t think I know those people personally. And to add a political dimension (heh), it’s relevant that most non-LW people (like me) initially encounter “politics is the mind-killer” being thrown out in comment threads, not through reading the original article. My opinion of the concept improved a lot once I read the article.

In the same thread, Andrew Mahone added, “Using it in that sneering way, Miri, seems just like a faux-rationalist version of ‘Oh, I don’t bother with *politics*.’ It’s just another way of looking down on any concerns larger than oneself as somehow dirty, only now, you know, *rationalist* dirty.” To which Miri replied: “Yeah, and what’s weird is that that *really* doesn’t seem to be Eliezer’s intent, judging by the eponymous article.”

Eliezer replied briefly, to clarify that he wasn't generally thinking of problems that can be directly addressed in local groups (but happen to be politically charged) as "politics":

> Hanson’s “Tug the Rope Sideways” principle, combined with the fact that large communities are hard to personally influence, explains a lot in practice about what I find suspicious about someone who claims that conventional national politics are the top priority to discuss. Obviously local community matters are exempt from that critique! I think if I’d substituted ‘national politics as seen on TV’ in a lot of the cases where I said ‘politics’ it would have more precisely conveyed what I was trying to say.

But that doesn't resolve the issue. Even if local politics is more instrumentally tractable, the worry about polarization and factionalization can still apply, and may still make it a poor epistemic training ground.

A subtler problem with banning “political” discussions on a blog or at a meet-up is that it’s hard to do fairly, because our snap judgments about what counts as “political” may themselves be affected by partisan divides. In many cases the status quo is thought of as apolitical, even though objections to the status quo are ‘political.’ (Shades of Pretending to be Wise.)

Because politics gets *personal* fast, it’s hard to talk about it successfully. But if you’re trying to build a community, build friendships, or build a movement, you can’t outlaw everything ‘personal.’

And *selectively* outlawing personal stuff gets even messier. Last year, daenerys shared anonymized stories from women, including several that discussed past experiences where the writer had been attacked or made to feel unsafe. If those discussions are made off-limits because they relate to gender and are therefore ‘political,’ some folks may take away the message that they aren’t allowed to talk about, e.g., some harmful or alienating norm they see at meet-ups. I haven’t seen enough discussions of this failure mode to feel super confident people know how to avoid it.

Since this is one of the LessWrong memes that’s most likely to pop up in cross-subcultural dialogues (along with the even more ripe-for-misinterpretation “policy debates should not appear one-sided“…), as a first (very small) step, my action proposal is to obsolete the ‘mind-killer’ framing. A better phrase for getting the same work done would be ‘**politics is hard mode**’:

1. ‘Politics is hard mode’ emphasizes that ‘mind-killing’ (= epistemic difficulty) is quantitative, not qualitative. Some things might instead fall under Middlingly Hard Mode, or under Nightmare Mode…

2. ‘Hard’ invites the question ‘hard for whom?’, more so than ‘mind-killer’ does. We’re used to the fact that some people and some contexts change what’s ‘hard’, so it’s a little less likely we’ll universally generalize.

3. ‘Mindkill’ connotes contamination, sickness, failure, weakness. In contrast, ‘Hard Mode’ doesn’t imply that a thing is low-status or unworthy. As a result, it’s less likely to create the impression (or reality) that LessWrongers or Effective Altruists dismiss out-of-hand the idea of hypothetical-political-intervention-that-isn’t-a-terrible-idea. Maybe some people do want to argue for the thesis that politics is always useless or icky, but if so it should be done in those terms, explicitly — not snuck in as a connotation.

4. ‘Hard Mode’ can’t readily be perceived as a personal attack. If you accuse someone of being ‘mindkilled’, with no context provided, that smacks of insult — you appear to be calling them stupid, irrational, deluded, or the like. If you tell someone they’re playing on ‘Hard Mode,’ that’s very nearly a compliment, which makes your advice that they change behaviors a lot likelier to go over well.

5. ‘Hard Mode’ doesn’t risk bringing to mind (e.g., gendered) stereotypes about communities of political activists being dumb, irrational, or overemotional.

6. ‘Hard Mode’ encourages a growth mindset. Maybe some topics are too hard to ever be discussed. Even so, ranking topics by difficulty encourages an approach where you try to do better, rather than *merely* withdrawing. It may be wise to eschew politics, but we should not *fear* it. (Fear is the mind-killer.)

7. **Edit:** One of the larger engines of conflict is that people are so much worse at noticing their own faults and biases than noticing others'. People will be relatively quick to dismiss others as 'mindkilled,' while frequently flinching away from or just-not-thinking 'maybe I'm a bit mindkilled about this.' Framing the problem as a challenge rather than as a failing might make it easier to be reflective and even-handed.

This is not an attempt to get more people to talk about politics. I think this is a better framing *whether or not* you trust others (or yourself) to have productive political conversations.

When I playtested this post, Ciphergoth raised the worry that 'hard mode' isn't scary-sounding enough. As dire warnings go, it's light-hearted—exciting, even. To which I say: *good*. Counter-intuitive fears should usually be *argued into* people (e.g., via Eliezer's politics sequence), not connotation-ninja'd or chanted at them. The cognitive content is more clearly conveyed by 'hard mode,' and if some group (people who love politics) stands to gain the most from internalizing this message, the message shouldn't cast that very group (people who love politics) in an obviously unflattering light. LW seems fairly memetically stable, so the main issue is what would make this meme infect friends and acquaintances who haven't read the sequences. (Or *Dune*.)

If you just want a scary personal mantra to remind yourself of the risks, I propose 'politics is SPIDERS'. Though 'politics is the mind-killer' is fine there too.

If you and your co-conversationalists haven’t yet built up a lot of trust and rapport, or if tempers are already flaring, conveying the message ‘I’m too rational to discuss politics’ or ‘You’re too irrational to discuss politics’ can make things worse. In that context, ‘politics is the mind-killer’ is the mind-killer. At least, it’s a needlessly mind-killing way of warning people about epistemic hazards.

‘Hard Mode’ lets you speak as the Humble Aspirant rather than the Aloof Superior. Strive to convey: ‘I’m worried I’m too low-level to participate in this discussion; could you have it somewhere else?’ Or: ‘Could we talk about something closer to Easy Mode, so we can level up together?’ More generally: If you’re worried that what you talk *about* will impact group epistemology, you should be even more worried about *how you talk about it*.

Change Contexts to Improve Arguments

On a recent trip to Ireland, I gave a talk on tactics for having better arguments (video here). There's plenty in the video that's been discussed on LW before (Ideological Turing Tests and other reframes), but I thought I'd highlight one other class of trick I use to have more fruitful disagreements.

It's hard, in the middle of a fight, to remember, recognize, and defuse common biases, rhetorical tricks, emotional triggers, etc. I'd rather cheat than solve a hard problem, so I put a lot of effort into shifting disagreements into environments where it's easier for me and my opposite-number to reason and argue well, instead of relying on willpower. Here's a recent example of the kind of shift I like to make:

A couple months ago, a group of my friends were fighting about the Brendan Eich resignation on facebook. The posts were showing up fast; everyone was, presumably, on the edge of their seats, fueled by adrenaline, and alone at their various computers. It’s a hard place to have a charitable, thoughtful debate.

I asked my friends (since they were mostly DC based) if they’d be amenable to pausing the conversation and picking it up in person. I wanted to make the conversation happen in person, not in front of an audience, and in a format that let people speak for longer and ask questions more easily. If so, I promised to bake cookies for the ultimate donnybrook.

My friends probably figured that I offered cookies as a bribe to get everyone to change venues, and they were partially right. But my cookies had another strategic purpose. When everyone arrived, I was still in the process of taking the cookies out of the oven, so I had to recruit everyone to help me out.

“Alice, can you pour milk for people?”

“Bob, could you pass out napkins?”

“Eve, can you greet people at the door while I’m stuck in the kitchen with potholders on?”

Before we could start arguing, people on both sides of the debate were working on taking care of each other and asking each others’ help. Then, once the logistics were set, we all broke bread (sorta) with each other and had a shared, pleasurable experience. *Then* we laid into each other.

Sharing a communal experience of mutual service didn’t make anyone pull their intellectual punches, but I think it made us more patient with each other and less anxiously fixated on defending ourselves. Sharing food and seating helped remind us of the relationships we enjoyed with each other, and why we cared about probing the ideas of this particular group of people.

I prefer to fight with people I respect, who I expect will fight in good faith. It's hard to remember that's what I'm doing if I argue with them in the same forums (comment threads, fb, etc) that I usually see bad fights. An environment shift and other compensatory gestures makes it easier to leave habituated errors and fears at the door.

Crossposted/adapted from my blog.