Posted by u/Prof_Nick_Bostrom Founder|Future of Humanity Institute 6 years ago

Science AMA Series: I'm Nick Bostrom, Director of the Future of Humanity Institute, and author of "Superintelligence: Paths, Dangers, Strategies", AMA

Superintelligence AMA

I am a professor in the faculty of philosophy at Oxford University and founding Director of the Future of Humanity Institute and of the Programme on the Impacts of Future Technology within the Oxford Martin School.

I have a background in physics, computational neuroscience, and mathematical logic as well as philosophy. My most recent book, Superintelligence: Paths, Dangers, Strategies, is now an NYT Science Bestseller.

I will be back at 2 pm EDT (6 pm UTC, 7 pm BST, 11 am PDT), Ask me anything about the future of humanity.

-- You can follow the Future of Humanity Institute on Twitter at @FHIOxford and The Conversation UK at @ConversationUK.

523 Comments Give Award Share ••• 85% Upvoted

This thread is archived
New comments cannot be posted and votes cannot be cast

SORT BY Best

View discussions in 1 other community

logos__ 172 points · 6 years ago

Professor Bostrom,

If a bear were to write a book about superbears, he would imagine them to be larger, faster, stronger, more powerful, have bigger claws, and so on. This is only natural; he doesn't have anything but himself to draw inspiration from. Consequently, he also would never be able to conceive of a human being, a being so much more in control of the world that we are both in complete control of its life and completely incomprehensible to it.

accurate, and not just a bear imagining a superbear? If the step from us to superintelligence is comparably transformative as the step from chimpanzee to us, how could we ever say anything sensible about it, being the proverbial chimpanzee? I imagine a chimpanzee philosopher thinking about superchimpanzees, and the unbelievably efficient and enormous ant siphoning sticks they would be able to develop, never realizing that, perhaps, the superchimpanzees would never even consider eating ants, let alone dream up better ant harvesting methods.

Share •••

Prof_Nick_Bostrom Founder|Future of Humanity Institute 119 points · 6 years ago

Yes, it's quite possible and even likely that our thoughts about superintelligences are very naive. But we've got to do the best we can with what we've got. We should just avoid being overconfident that we know the answers. We should also bear it in mind when we are designing our superintelligence - we would want to avoid locking in all our current misconceptions and our presumably highly blinkered understanding of our potential for realizing value. Preserving the possibility for "moral growth" is one of the core challenges in finding a satisfactory solution to the control problem.

Share •••

jinxr 25 points · 6 years ago

Ha, "bear it in mind", I see what you did there.

Share •••

1 more reply

2 more replies

Smallpaul 14 points · 6 years ago

You might be right.

But in a recent discussion among scientists and philosophers one of them made the point that this analogy is a bit weird. A bear can't imagine a super-bear because a bear can't reason. A chimp can't imagine a super-chimp because a chimp does not have that imaginative potential.

There are all kinds of reasons that we might (almost certainly are!) wrong about our imaginings of superintelligences, but the delta between us and them may or may not be one. A much simpler example might be Alexander Graham Bell trying to imagine a "smartphone". Cognitive differential is not necessarily the problem.

Share •••

logos__ 15 points · 6 years ago

That is the exact issue. Among living things, cognition is a scale. Compared to bacteria, bears are smart; they can evade predators, seek out food, store it, and so on. Compared to us, bears are dumb. They can't talk, they can't pay with credit cards, they can't even play poker. At some points on that scale, small incremental

one between us and dolphins (and every other form of life). There's also one between us and superintelligences. Our cognition allows up to see the next qualitative bump up (whereas this is denied to, say, a chimpanzee), but it doesn't allow us to see over it. That's the problem.

Share •••

lheritier1789 BS | Chemistry Psychology 2 points · 6 years ago

It seems like we don't necessarily need to see over it. Can we not evolve in a stepwise fashion, where each iteration conceives of a better version?

It seems totally plausible that a chimp might think, hey, I'd like to learn to use these tools faster. And if he were to have some kind of method to progress in that direction, then after some number of iterations you might get a more cognitively developed animal. And it isn't like the initial chimp has to already know that they were going to invent language or do philosophy down the line. They would just need higher computing power and complex reason seems like it could conceivably arise that way.

So I don't think we have to start with some kind of ultimate being. We just have to take it one step at a time. We'll be a different kind of being once we get to our next intelligence milestone, and those beings will figure out their next steps themselves.

Share •••

dalabean 5 points · 6 years ago

The issue is with a self improving super-intelligence those steps could happen a lot faster than we have time to understand what is happening.

Share •••

FlutterNickname 2 points · 6 years ago

All that will matter is that the super intelligences understand it. They would no more want to defer decisions to us than we would to the bear.

Therein lies the potential need for transhumanism.

Imagine a world where the super intelligences already exist and have become commonplace. Keeping up as an individual, if desired, means augmentation of some sort. At a cognitive level, normal humans will be just another lower primate, and we'll be somewhat dependent on their altruism.

Share •••

1 more reply

1 more reply

JazzerciseMaster 17 points · 6 years ago

Where would one find these super bears? Is this something we should be worried about?

tilkau 18 points · 6 years ago

Don't be silly. Super bears find *you*.

Share •••

TheNextWhiskyBar 5 points · 6 years ago

Not if you pay the Bear Patrol tax.

Share •••

[deleted] 2 points · 6 years ago

No youll be fine. As long as its not a super seabear and you arent wearing a sombrero wrong.

Share •••

Ungrateful_bipedal 2 points · 6 years ago

I just laughed so hard I nearly woke up my son. Imaginary gold for you sir.

Share •••

1 more reply

1 more reply

categorygirl 2 points · 6 years ago

Chimp's can't even linearly extrapolate like the way we do it. People 5000 years ago imagined flying machines. Humans have figured out physics so we can use physics to constraint what is possible. We may not be able to linear extrapolate but we could still hit a possible good guess (chimps won't even make a good guess about a space elevator stick). But I also think your example could be true too. Maybe our understanding of physics is like the chimps understanding of the stick.

Share •••

15 more replies

Comment deleted by user 6 years ago 6 children

404random 40 points · 6 years ago · *edited 6 years ago*

Hi Dr Bostrom, As a debater we use a lot of your work to talk about extinction. I have two questions. The first is what do you think is the most likely threat of extinction in the coming century? Is it a natural impact or is it war? The second is that in 2001 you wrote an article saying that US Russia war is the most likely war scenario for extinction. I believe in 2007 you wrote another article which talked about how Russian war will not cause extinction. Which statement do you agree with and what war most likely will cause extinction? Also can I quote you on your answers here? Thank you Edit: Have you read any work by Marshall Savage and if you have do you agree with any of his work?

Share •••

There are two questions we must distinguish: what is the biggest existential risk right now, and what is the biggest existential risk. Conditional on something destroying us in the next few years, maybe nuclear war and nuclear winter are high on the list (even though our best bet is that they wouldn't cause our extinction even if they occurred). But I think there will be much larger xrisks in the future - risks that are basically zero today (e.g. from superintelligence, advanced synthetic biology, nanotech, etc.)

Not familiar with work of Savage. (Feel free to quote me there, but don't quote me when I say that continental philosophy in college debating is a worrisome source of xrisk...)

Share •••

Comment deleted by user 6 years ago 2 children

1 more reply

coherent_sheaf 11 points · 6 years ago

> The second is that in 2001 you wrote an article saying that US Russia war is the most likely war scenario for extinction. I believe in 2007 you wrote another article which talked about how Russian war will not cause extinction.

Those statements don't contradict each other. To compare: the most likely way for me to die today is to get hit by a car when I go grocery shopping; I will probably not get hit by a car when I go to the grocery.

Share •••

3 more replies

3 more replies

punctured-torus 25 points · 6 years ago · *edited 6 years ago*

Hi Dr. Bostrom,

- When you discuss "infrastructure profusion" you highlighted some negative unintended consequences of AI utilizing the solar system as a "computronium" to solve complex mathematical problems. What are some other unintended consequences that you foresee (not highlighted in your book)?
- What are some examples of problems that you consider *robustly positive* and *robustly justifiable*?
- Do you feel like AI can be achieved without consciousness? Do you feel like the two are intrinsically connected? *Disclaimer: Whatever consciousness means.*
- In your opinion, do you feel like the rewards reaped from achieving AI outweighs the risk of it?

Share •••

Prof_Nick_Bostrom Founder|Future of Humanity Institute 15 points · 6 years ago

arising from coordination failures in multipolar outcomes.)

2. It's a matter of degree - it's surprisingly hard to think of any problem that is extremely robustly positive to the extent that we can be fully certain that a solution to it would be on balance good. But, for example, making people kinder, increasing collective wisdom, or developing better ways to promote world peace, collaboration, and compromise seem fairly robustly positive.
3. I don't feel I understand the exact computational prerequisites for consciousness well enough to have a strong view on that.
4. These kinds of question I think need to be answered relative to some alternative, and it is not clear in this case what the alternative is relative to which achieving AI would or would not be better. But if the question is, would it be good or bad news if we somehow discovered that it is physically impossible ever to create superintelligence, then the answer would seem to be that it would be bad news.

Share •••

2 more replies

jumbowumbo 52 points · 6 years ago

I'm the head of the Futurism Society at Tufts University. I attended your recent talk at Harvard and I never got to ask my question there.

If I can ask you to be self-critical here, are there any reasons you can think of to be skeptical of investing our time and energy into mitigating existential risk? The concept seems awfully close to Nozick's utility demon.

Share •••

Prof_Nick_Bostrom Founder|Future of Humanity Institute 49 points · 6 years ago

One worry is that the study of xrisk could generate information hazards that lead to a net increase in xrisk.

From a moral point of view, it's possible that aggregative ethics is false; and that some other ethical theory is true that would imply that preventing extinction is much less important.

I've written about the problems aggregative consequentialism faces when one considers the possibility of infinite goods - it threatens ethical paralysis, which could suggest it is always morally indifferent what we do.

From a selfish point of view, the the level of xrisk may be low enough that it is not a dominant concern, and hard enough to influence that it wouldn't warrant investing any resources.

Share •••

narwi 3 points · 6 years ago

So essentially, by studying xrisk we make xrisk actualising more likely, as people would seek to weaponise it for yet another mutually assured destruction scenario?

Share •••

scholl_adam 4 points · 6 years ago

If another ethical theory were true -- non-cognitivism, say -- that could be a huge risk itself, right? If a superintelligence discovers that the moral system we've imbued it with is flawed, it would be rational for it to adopt one that corresponds more closely with reality... and we might not like the results.

Share •••

FeepingCreature 6 points · 6 years ago

Ethics relates to utility. What's ethical is not the same kind of question as what's true. If I have a preference for ice cream, this describes reality only insofar as this fact is part of the physical makeup of my brain. To the best of my understanding, an ethical claim cannot be true or untrue. - I'm trying to think of examples, but all the ethical statements I can think of are in fact more like truths about my brain. Which of course can be wrong - I might simply be wrong about my own preferences. But I don't see how *preferences*, per se, can be wrong; even though every sentence I could use to communicate them can be.

AFAICT, The only way we could get problems with truths or untruths in ethics, is if the *description* of ethical preferences that the AI works on is inconsistent or flawed.

Share •••

scholl_adam 7 points · 6 years ago

I agree with you; A.J. Ayer and many others would too. But there are also a lot of folks (moral realists) who disagree. My point was just that it makes saftey-sense for AI researchers to assume that their ethical frameworks -- no matter how seemingly-desirable -- are not literally true *even if* they are committed moral realists. When programming a superintelligent AI, metaethical overconfidence could be extremely dangerous.

Share •••

1 more reply

easwaran 2 points · 6 years ago

That's a controversial meta-ethical view. It strikes me that some sort of moral realism is more plausible. I agree that moral facts seem like weird spooky facts, but I think they're no more spooky than other facts that we all do accept.

Presumably you think it's correct to say that evolution is a better justified theory of the origin of species than creationism. Furthermore, evolution is a better justified theory now than it was in 1800. And there might be other things that we're justified in believing given our current evidence, even though they turn out not to in fact be true.

latter is too spooky to accept, then I'm not quite sure how you save the former. And to deny that one belief is ever better justified than another seems to me to involve giving up a whole lot.

Share •••

1 more reply

1 more reply

1 more reply

nallen PhD | Organic Chemistry 48 points · 6 years ago

Science AMAs are posted early, with the AMA starting later in the day to give readers a chance to ask questions vote on the questions of others before the AMA starts.

Prof. Bostrom is a guest of r/science and has volunteered to answer questions. Please treat him with due respect. Comment rules will be strictly enforced, and uncivil behavior will result in a loss of privileges in r/science.

if you have scientific expertise, please verify this with our moderators by getting your account flaired with the appropriate title. Instructions for obtaining flair are here: reddit Science Flair Instructions

Flair is automatically synced with r/EverythingScience as well.

Share •••

Eight_Rounds_Rapid 59 points · 6 years ago

Good evening from Australia Professor! I would really like to know what your opinion is on **technological unemployment**. There is a bit of a shift in public thought and awareness at the moment about the rapid advances in both software and hardware displacing human workers in numerous fields.

Do you believe this time is actually different compared to the past and we do have to worry about the economic effects of technology, and more specifically AI, in permanently displacing humans?

Thanks!

Share •••

Prof_Nick_Bostrom Founder|Future of Humanity Institute 80 points · 6 years ago

It's striking that so far we're mainly used our higher productivity to consume more stuff rather than to enjoy more leisure. Unemployment is partly about lack of income (fundamentally a distributional problem) but it is also about a lack of self-respect and social status.

I think eventually we will have technological unemployment, when it becomes cheaper to do most everything humans do with machines instead. Then we can't make a living

us cultivate interest in activities that are not done to earn money.

Share •••

davidmanheim 7 points · 6 years ago

Is it only a cultural sigma that surrounds idleness? Many studies seem to show that people are dissatisfied without something they view as productive work.

The idea that we can transition to a culture where the sigma is gone ignores this important question - and the outcome may argue for strictly limiting the power of computers and machine learning systems, instead of attempting to keep them benevolent, which may not be possible. (Coordination problems may make this an unfeasible solution, though.)

Share •••

Smallpaul 15 points · 6 years ago

> Is it only a cultural sigma that surrounds idleness? Many studies seem to show that people are dissatisfied without something they view as productive work.

There is a lot of knitting, painting, singing, composing, gardening, rainbow looming, electronics hacking and writing to be done.

People still get very emotionally attached to amazing Chess games:

- http://theamazingchessworld.blogspot.ca/

Would you say that very "talented" chess pros are just wasting their lives because a computer could "do it better"? Do they lack self-worth and life satisfaction?

Share •••

3 more replies

saibog38 5 points · 6 years ago

> Many studies seem to show that people are dissatisfied without something they view as productive work.

This is only really an issue if you define "productive work" as that which produces monetary value. At least for me, the majority of my most satisfying endeavors are those that don't directly produce any monetary value, but are nonetheless deeply satisfying (you could even say priceless) to myself.

Share •••

14 more replies

bushwakko 2 points · 6 years ago · *edited 6 years ago*

You are conflating paid work or jobs with actual doing labor or work. Even if you cannot get a job at McDonald's (which people usually don't find all that fulfilling anyway) working on your home, raising kids, getting a hobby etc are all things

Edit: also, one reason that jobless people cannot find satisfying things to do at the moment is that they literally aren't allowed to do productive things like start their own business etc because a condition for getting welfare is basically that you cannot do that.

Share •••

2 more replies

someguyfromtheuk 21 points · 6 years ago

As a member of the public, it seems like this time is worse.

Before, if you replaced a job with a technology, that technology was still made by and repaired by other human beings, so there were jobs being created.

If we build an AI capable of doing anything a human can do, or a robot capable of any physical movement a human is, then they can effectively replace all jobs, since any new job created for humans, could be done by the robots, and probably faster since they don't need to eat or sleep.

Thus, it seems like a permanent change, and one that modern society doesn't really seem equipped to deal with, a lot of people still have the attitude that a person's value is based around how much he works, and that people only deserve things if they work for them, which don't fit into a society where 99% of human workers are replaced by AI or robots.

Share •••

MaeveSuave 22 points · 6 years ago

"It seems like this time is worse."

I hear that sentiment concerning jobs, and it's a strange thing. Here we have, for all intents and purposes, this situation: "technology is doing the work of more men. Where once 10 were needed, now only 2 are needed to do the same thing." And this means that, in the case of agriculture for example, 2 people provide the same amount of food as 10 once did. Step outside the economic structure we've created, see the abundance in every grocery store, see the free time that is thereby created, and well... by all objective standards, during a time of abundance, unemployment, here, is a good thing.

Question is, now, how do we adjust our economic framework to utilize that as best as possible? Because if we can, I think we're talking about a new renaissance here.

Share •••

someguyfromtheuk 16 points · 6 years ago

Yes, I understand that this could be a major turning point for the better, a time free of scarcity, but frankly, our economy still requires money to buy things, and completely dismantling that would be the reversal of tens of thousands of years of

If we don't move into a more socialist form of society, then inequality will keep rising and rising until society collapses because it's simply unsustainable.

Share •••

Herculius 13 points · 6 years ago · *edited 6 years ago*

As much as Marx's ideas have gone out of fashion I think his materialistic conception of the means of production will be useful. As it stands the productivity and efficiency increases of computers and machines serve the owners of the means of production. Corporations and businesses use patents and barriers to entry to decrease costs and improve the utility of their products.

In this environment people in control of productive assets are becoming less dependant on labor and the general public. The corrallary is that the general public is becoming more dependant on productive assets controlled by those with ownership.

What I'm attempting to get at is that we need a different way to think about ownership and control of hardware and software so that technology works for individuals and not just the elite.

People need to realise how much power and knowledge is already at their fingertips and fight tooth and nail to make sure technology is working for them.

A few examples of how technology could empower individuals are: *more widespread 3d printers to create and modify our own tools, *open source software/hardware so that you are free to improve and modify the technology you use, and the *freedom of information and education so that low and middle class individuals aren't excluded from the technical ways and they could increase their own autonomy

The powers that be think they know what's best for you and your future, and they want you to trust and believe them. And if you don't comply they will use pre-existing legal structures to make sure they maintain control.

I hope this isn't too much conjecture for r/science but the futuristic and political topic seems like it would benefit from varying perspectives.

Share •••

1 more reply

Orwelian84 11 points · 6 years ago

It doesn't even have to get to the 99% level to be "catastrophic" from a societal standpoint. The great recession and depression were both below 30% unemployment and they were definitely difficult for society to deal with.

5%ish unemployment(thank you Milton Friedman).

Imagine if we have to reorganize around 10-15% unemployment being the structural baseline. That doesn't require super intelligent AI, just the deployment and scaling of existing programs like Watson and partial automation of the transportation industry.

Share •••

someguyfromtheuk 4 points · 6 years ago

Yeah, I know it doesn't need to be 99%, that was just an extreme example.

Yeah, I'm with you on unemployment hitting difficult levels relatively soon, self driving vehicles could automate away a lot of jobs like taxi driver, bus driver, pilot, train driver, ship pilot etc. and then there's self serving kiosks eliminating cashiers, AI decreasing the amount of middle management required and just the general increase in productivity due to technology meaning a drop in the number of workers required for pretty much anything.

I think the last things to be automated will be manual jobs like construction or loading/unloading vehicles or waitering, along with creative jobs like artists and scientific innovation, although technology can make them more productive, so there'd be less of them.

Frankly, I think more individualistic countries like America are going to end up worse than countries with a more socialistic mindset like Scandinavian countries or East Asian ones, since it'll be harder for them to implement the wide scale social programs that'll be needed like Basic Income and socialised healthcare and education.

Share •••

Orwelian84 7 points · 6 years ago

I tend to agree, although America does have a history of coming together, we just take our sweet time getting around to it.

Any job, regardless of the field, that can be brute forced(in the software sense) is liable to be replaced by automation over the next decade i think.

I can imagine an American population getting behind the idea of a Negative Income Tax as a form of Basic Income, but it will take the beginning of the die off of the boomers for it to be politically viable. Too much fear of "socialism" and "communism" left in that generation from the Red Scare.

Share •••

[deleted] 4 points · 6 years ago

Next decade? Probably not. Eventually yes, but you have to remember that any means of trying to supplant a large section of the workforce takes time and will be meet with resistance. The general phase out of domestic customer service employees serves as a decent model. The means (foreign

occurred.

Share •••

Orwelian84 3 points · 6 years ago

I totally agree, I say within a decade because the automation won't be heavily focused on any one industry(the transportation industry aside), but rather on most of them. Even if it is half a percent every five years if that half a percent comes out of every single industry the net effect could be like I fear, 10-15% structural unemployment by 2025.

I don't doubt there will be resistance, I am just not sure how we could do anything about it. If we don't automate our "rivals" will, we are caught between a rock and a hard-place.

Share •••

[deleted] 4 points · 6 years ago

Yeah it will be interesting. I've often thought about how difficult it will be to explain to the millions of America's truck drivers that a computer can get the load to the client faster and safer while using less fuel.

Share •••

2 more replies

1 more reply

1 more reply

bertbarndoor 2 points · 6 years ago

If inputs resource scarcity is eliminated, then robotic replacements (AI not essential) can fulfill humanities survival-dependent heirarchy of needs (food/shelter). This will redefine the meaning of value, wealth, and class. Imagine a nearly perfectly efficient post energy-grid-parity world where all material physical inputs into any prodcution system are sourced, maipulated, and delivered to end users by mechanical means without human intervention.

Share •••

5 more replies

1 more reply

jmdugan PhD | Biomedical Informatics | Data Science 11 points · 6 years ago

Do you believe there is something mystical or undefinable about human consciousness, or do you believe it is a series of explainable functions that we can map out and understand, or possibly something else?

Share •••

of sensory inputs and corresponding brain modules. Ask yourself, why would a creature such as ourselves, that is evolved to see, hear, taste, touch and store those senses and process those senses NOT do those things? People ask, "yes, but WHY are we conscious?!?!" Think about what those people are actually asking - really think about it. Understand what consciousness really is - sensory inputs and corresponding processes. Consciousness is simply the functioning of those process. So why WOULDN'T your eyes see? Why WOULDN'T your brain observe those visuals from a fixed perspective, etc... etc... There is no mystery here. There is no "hard problem". The actual hard problem is understanding how those senses and corresponding processes actually work - how the cells function and are connected at a fundamental level- not in realizing that those process really do what they were evolved to do.

Share •••

ihaveahadron 2 points · 6 years ago · *edited 6 years ago*

I agree with what you have to say. I think philosophers are morons. However, I have one question about the subject--without using bullshit terminology like a "hard problem".

I understand *why* our bodies react the way they do--due to brain interactions. But why is it that we *experience* those interactions? It has been fully explained to me why everything in human history has happened--including the existence of all life, and what it has done. However, I don't see the explanation as to why all of the organisms are able to "feel" and "experience" the senses which are created in their brains.

It seems plausible to me that all of life and it's actions could have taken place--but yet none of it's members could have ever been aware of it.

I understand that the fact that because we are each indivually able to experience the feelings created in our brains, that the latter scenario is proven to not be possible--however, is there a scientific answer that could explain the phenomena of concioussness?

And a further question is--do computer circuits experience some form of concioussness? If not, what makes them different from organic forms of circuitry?

Share •••

daerogami 3 points · 6 years ago · *edited 6 years ago*

That's a great set of questions. I would like to provide some input on the last two questions (answering the last should address the first). I spent a fair amount of time studying Neural Networks while at university. While I wouldn't say this makes me qualified to provide a perfect or scientifically acceptable answer, I hope just the same it sheds some light on the topic.

Computer circuits are made up of what are called gates (the most fundamental level that computers "process"). And these gates take input and provide output. These properties are important to note:

- The gates always processes the same exact way, every time. It is a static function.
- The connections to these gates are also static. The source always comes from the same gates preceding it and to output always goes to the gate following it.

In order to stick within the boundaries of my knowledge I will give the computational corollary to the human brain, a neural network. Neural networks are modelled after the organic brain; what is known as 'biologically inspired'. Their most fundamental level of processing are neurons. Much like gates, they take input and provide output. The following points are respective to the preceding points:

- Neurons input and output can span a wide range of 'signals' (such as all integers), the human brain, IIRC, has 7 different chemical signals (known as neurotransmitters).
- Neurons can 'learn' from previous input and retrieve feedback from other neurons which allows them to modify the way they process input.
- The most mind boggling part of the organic brain (at least to me) is that the neurons can change their connections with other neurons. I don't understand exactly how it works, but I have not heard of a neural network that simulates this. To feed your curiosity if you wish to dig further

I hope this has brought some insight into the 'conciousness' of computers vs brains. Again, please note, I am not an authority on this material and it may very likely contain inaccuracies.

Share •••

ihaveahadron 2 points · 6 years ago

Thanks for the reply. That is new information to me.

Share •••

3 more replies

1 more reply

tygrgo 78 points · 6 years ago · *edited 6 years ago*

Hi, Professor Bostrom.

In the current issue of The New York Review of Books, John Searle subjects your book/thinking to quite a take down: "I believe that neither [your book nor Luciano Floridi's] gives a remotely realistic appraisal of the situation we are in with computation and information." He writes that the reason for this, "in its simplest form, is that they fail to distinguish between the real, intrinsic observer-independent phenomena corresponding to these words and the observer-relative phenomena that also correspond to these words but are created by human consciousness."

Do you feel that Searle accurately represented your position in his article? Eager to know if you have any plans to respond to Searle's article, and if you might lay out some of what you would want to include in such a response here today. Thank you so much!

[EDIT] - fixed typo

Share •••

Prof_Nick_Bostrom Founder|Future of Humanity Institute 66 points · 6 years ago

The answer to your question is no. For example, Searle seems to think that I'm convinced that superintelligence is just around the corner, whereas in fact I'm fairly agnostic about the time frame.

Obviously I also have more substantial disagreements with his views. I disagree with him about the metaphysics of mind and with the implications he wants to draw from his Chinese room thought experiment. I think he has been refuted many times over by lots of philosophers, and I don't feel the need to go over that again. But the disagreement seems to extend beyond the metaphysical question of whether computers could be conscious. He seems to say that computers don't "really" compute, and that therefore superintelligent computers would not "really" be intelligent. And I say that however that might be, they could still be dangerous. (And dead really is dead.)

Share •••

jenkc 14 points · 6 years ago

link, paywalled :(http://www.nybooks.com/articles/archives/2014/oct/09/what-your-computer-cant-know/

Share •••

[deleted] 44 points · 6 years ago

i read a recent interview with Searle and he still doesn't seem to understand the flaw in his chinese room. i'd like to give him the respect he's due... but at some point you have to realize that you're arguing with (effectively) a creationist that is so emotionally entrenched in his positions that all he has left is angry barely coherent rants. when you try to politely ignore him, he declares victory.

Share •••

[deleted] 16 points · 6 years ago

For those of us unfamiliar with this subject basically at all, would you care to enlighten us? Because at present you're just saying "he doesn't see how wrong he is, duh" which of course to the uninformed observer is not helpful.

Share •••

dragonnyxx 54 points · 6 years ago · *edited 6 years ago*

The "Chinese room" is a thought experiment he proposed. Imagine a room containing an arbitrary number of filing cabinets full of arbitrarily complicated

cabinets to (in some way) "process" the symbols on the sheet of paper and compose a reply, again consisting of some sorts of symbols. We allow him arbitrary time to finish the response and assume he will never make a mistake. He places this reply in the out-box. Because he's just following the instructions, he doesn't actually understand what the symbols mean.

Unbeknownst to the person in the room, the symbols he is processing are Chinese sentences, and the responses he is producing (by following these arbitrarily complicated instructions) are also Chinese sentences -- responses to the input. The filing cabinets contain, essentially, a computer program smart enough to understand Chinese text and respond appropriately, as a human would, and the person in the room is essentially "running the program" by virtue of following the instructions. The room can "learn" via instructions commanding the person to write things down, update instructions and so forth, so it can be a perfectly good simulation of a Chinese-speaking person.

Ok, fine.

Now, Searle argues that because the person in the room doesn't actually *understand* Chinese, that computers can't really "understand" things in the way we do and thus computers cannot really be intelligent.

This is, of course, a completely asinine argument. It's true that one small part of the overall system -- the person (equivalent to the computer's processor) -- does not actually understand Chinese, but *the system as a whole* certainly does. But basically Searle is a master of ignoring perfectly good arguments, deflecting, and moving the goalposts, so he will never at any point admit that it is possible for something other than a human brain to really "understand" something.

The more astute folks in the audience will of course note that we don't actually have a good definition of what it means to really "understand" something (for instance, your computer can almost certainly perform math better than you can -- but does it really "understand" math?) I don't believe Searle provides a solid definition of this either; he basically just implicitly treats "understand" as "something humans do and computers don't", and then acts surprised when he reaches the conclusion that computers can't actually understand things.

Share •••

wokeupabug 39 points · 6 years ago · *edited 6 years ago*

Here's how you characterize Searle's position:

> But basically Searle is a master of ignoring perfectly good arguments, deflecting, and moving the goalposts, so he will never at any point admit that it is possible for something other than a human brain to really "understand" something.

This is a pretty common characterization of his position, which one can find pretty ubiquitously on internet forums whenever his name pops up.

Here's what Searle actually writes in the very article you were commenting on:

> For clarity I will try to [state some general philosophical points] in a question and answer format, and I begin with that old chestnut of a question: "Could a machine think?" The answer is, obviously, yes. We are precisely such machines. "Yes, but could an artifact, a man-made machine think?" Assuming it is possible to produce artificially a machine with a nervous system, neurons with axons and dendrites, and all the rest of it, sufficiently like ours, again the answer seems to be obviously, yes. If you can duplicate the causes, you can duplicate the effects. And indeed it might be possible to produce consciousness, intentionality, and all the rest of it using some other sort of chemical principles than those human beings use. It is, as I [previously] said, an empirical question. "Ok, but could a digital computer think?" If by "digital computer" we mean anything at all that has a level of description where it can correctly be described as the instantiation of a computer program, then again the answer is, of course, yes, since we are the instantiations of any number of computer programs, and we can think. (Searle, "Minds, brains, and programs" in *Behavioral and Brain Sciences* 3:422)

I hope you can understand why my initial reaction, whenever I encounter the sort of *common wisdom* about Searle like that found in your comment, is to wonder whether the writer in question has actually read the material they're informing people about.

Readers of the article in question will recognize the objection you raise...

> This is, of course, a completely asinine argument. It's true that one small part of the overall system -- the person (equivalent to the computer's processor) -- does not actually understand Chinese, but the system as a whole certainly does.

... as being famously raised by... Searle himself in the very same article (p. 419-420).

It doesn't seem to me that it's particularly good evidence that Searle is "a master of ignoring perfectly good arguments" to point out an objection that he himself published. But if his article is to be credibly characterized as "completely asinine" by virtue of this objection, I would have expected you to have noted that he himself remarks upon this objection, and rebutted his objections to it.

Share •••

daermonn 7 points · 6 years ago

So what exactly is Searle's argument? Can you elaborate for us?

Share •••

timothymicah 4 points · 6 years ago

Searle's argument in a nutshell is that we KNOW that brains are sufficient for consciousness, but we don't know which elements are

the brain, it would almost certainly be conscious, but we wouldn't know why other than the fact that brains are sufficient for consciousness.

Furthermore, the Chinese Room argument is actually not a comment on artificial intelligence so much as a comment on the nature of intelligence itself. Minds, as we experience them, have semantic, meaningful contents. Computer programs consist of little more than syntactical structures, structures that do not contain inherently meaningful contents. Therefore, computer programs alone do not constitute minds. The mind is a semantic process above and beyond mere syntax.

Share •••

wokeupabug 2 points · 6 years ago

> Furthermore, the Chinese Room argument is actually not a comment on artificial intelligence so much as a comment on the nature of intelligence itself.

It is this, but it's also a comment not on artificial intelligence generally, but on a specific research project for artificial intelligence which was popular at the time.

> Searle's argument in a nutshell is that we KNOW that brains are sufficient for consciousness...

Right, so this is one of the differences: on Searle's view, neuroscience and psychology are going to make essential contributions to any project for AI, while proponents of the view he is criticizing often saw the specifics of neuroscience and psychology as fairly dispensable when it comes to understanding intelligence.

> Minds, as we experience them, have semantic, meaningful contents. Computer programs consist of little more than syntactical structures...

Right, this is the main thing in this particular paper. There's a question here regarding what's involved in intelligence, and on Searle's view there's more involved in it than is supposed by the view he's criticizing. In particular, as you say, Searle maintains that there is more to intelligence than syntactic processing.

This particular intervention into the AI debate might be fruitfully compared to that of Dreyfus, who likewise elaborates a critique of the overly formalistic conception of intelligence assumed by the classical program for AI. If we take these sorts of interventions seriously, we'd be inclined to push research into AI, or intelligence generally, away from computation in purely syntactical structures and start researching the way relations between organisms or machines and their environments produce the conditions for a semantics. And this is a lesson that the cognitive science community has largely taken to

Share •••

1 more reply

Incepticons 4 points · 6 years ago

Seriously thank you, its amazing how many people repeat the same "obvious flaws" in Searle's reasoning without ever reading...Searle.

The Chinese Room isn't bulletproof but wow is it attractive bait for people on here to show how philosophy is just "semantics"

Share •••

24 more replies

[deleted] 14 points · 6 years ago

Right. You could just as easily isolate cortices (cortexes?) in the brain and point out that there isn't evidence that the prefrontal cortex understands anything by itself or the visual cortex sees anything. The only important question is if the system as a whole does.

Share •••

Epistaxis PhD|Genetics 19 points · 6 years ago

It sounds like Searle is just using a roundabout scenario full of tempting distractions to camouflage the lack of a precise definition for *understand*, which is the main problem in the first place.

Share •••

Lujors 11 points · 6 years ago

Yes. Semantics.

Share •••

timothymicah 2 points · 6 years ago

Searle's argument in a nutshell is that we KNOW that brains are sufficient for consciousness, but we don't know which elements are necessary for consciousness. As a result, we're not sure how to begin building a conscious machine. If we built a machine that was identical to the brain, it would almost certainly be conscious, but we wouldn't know why other than the fact that brains are sufficient for consciousness. Furthermore, the Chinese Room argument is actually not a comment on artificial intelligence so much as a comment on the nature of intelligence itself. Minds, as we experience them, have semantic, meaningful contents. Computer programs consist of little more than syntactical structures, structures that do not contain inherently meaningful contents. Therefore, computer programs alone do not constitute minds. The mind is a semantic process above and beyond mere syntax.

21 more replies

[deleted] 1 point · 6 years ago

Great reply, thanks. (The instruction cards told me to say that).

I asked similar elsewhere: does this line of thinking spawn the Turing test? So a clever enough cleverbot can persuade you or I that it's human, do we declare that it understands?

As you mention the meaning of "understand" is really a fascinating question. Is the Chinese box "system" required to be able to provide a meaningful response, or does it simply provide a "satisfactory" response? That would seem essential to understanding the argument.

Share •••

techumenical 10 points · 6 years ago

It's probably best to see Searle's line of thinking as a counterargument to the idea underlying the Turing test--that is, all that is needed for a computer to be considered intelligent is that it is reasonably indistinguishable from a human in it's ability to converse. Searle would say that a computer system that passes the Turing test understands nothing and is therefore no more intelligent than a computer that can't pass the test.

The meaningfulness of the Chinese Room's response is "built" into the instructions provided to the room that the person follows when responding to inputs and, of course, in the interpretation of the response by those outsiders interacting with it. A more "meaningful" response could always be arbitrarily generated by updating the rules the person follows when processing inputs. The thrust of the Chinese Room argument is that the only possible thing to which we could attribute understanding, the human, is nothing more than a symbol processor. The meaningfulness of the responses is outside of the human's grasp since this human doesn't speak or recognize chinese. Therefore, nothing about the room can be said to understand anything.

Now, you might bring up the objection that the rules themselves constitute an understanding since they are the mechanism by which a "proper" response is generated, but that's a different post...

Share •••

dragonnyxx 3 points · 6 years ago

> The thrust of the Chinese Room argument is that the only possible thing to which we could attribute understanding, the human, is nothing more than a symbol processor. The meaningfulness of the responses is outside of the human's grasp since this human doesn't speak or recognize chinese. Therefore, nothing about the room can be said to understand anything.

more than relatively simple chemical switches, nothing about your brain can be said to understand anything.

Furthermore, "only possible thing to which we could attribute understanding, the human" is begging the question -- you are *assuming* that the human is the only thing capable of understanding. When you assume the conclusion your argument, it's little surprise when you reach that conclusion.

Share •••

techumenical 9 points · 6 years ago

It might be helpful to clarify that this is just my reading of the argument and that I provided it to help clarify some questions about "meaningfulness" and that concept's place in the discussion between Searle and Turing.

I would further mention that my reading is probably influenced by my belief that the Chinese Room Argument is flawed, so you may be noticing errors in my representation and not the argument itself.

I'd be happy to play devil's advocate to your points if there's interest, but I have the feeling that that's sort of beside the point here.

Share •••

Kuris 5 points · 6 years ago

This is the absolute classiest response to a potentially inflammatory post that I've ever seen!

No charged language or implications, and yet your point comes across perfectly!

Share •••

HabeusCuppus 2 points · 6 years ago

The Turing test is different and arguably spawned from things alan turing might have seen such as mechanical Turks.

Turing is more about whether or not an observer can distinguish and not whether a program is smart, anyway. And it's horribly calibrated

Share •••

1 more reply