Science AMA Series: I'm Joanna Bryson, a Professor in Artificial (and Natural) Intelligence. I am being consulted by several governments on AI ethics, particularly on the obligations of AI developers towards AI and society. I'd love to talk – AMA!

Joanna$_{B}ryson$[1]$andr/ScienceAMAs$[1]

[1]Affiliation not available

April 17, 2023

Abstract

Hi Reddit! I really do build intelligent systems. I worked as a programmer in the 1980s but got three graduate degrees (in AI & Psychology from Edinburgh and MIT) in the 1990s. I myself mostly use AI to build models for understanding human behavior, but my students use it for building robots and game AI and I've done that myself in the past. But while I was doing my PhD I noticed people were way too eager to say that a robot – just because it was shaped like a human – must be owed human obligations. This is basically nuts; people think it's about the intelligence, but smart phones are smarter than the vast majority of robots and no one thinks they are people. I am now consulting for IEEE, the European Parliament and the OECD about AI and human society, particularly the economy. I'm happy to talk to you about anything to do with the science, (systems) engineering (not the math :-), and especially the ethics of AI. I'm a professor, I like to teach. But even more importantly I need to learn from you want your concerns are and which of my arguments make any sense to you. And of course I love learning anything I don't already know about AI and society! So let's talk... I will be back at 3 pm ET to answer your questions, ask me anything!

the WINNOWER

REDDIT

Science AMA Series: I'm Joanna Bryson, a Professor in Artificial (and Natural) Intelligence. I am being consulted by several governments on AI ethics, particularly on the obligations of AI developers towards AI and society. I'd love to talk – AMA!

JOANNA_BRYSON R/SCIENCE

Hi Reddit!
I really do build intelligent systems. I worked as a programmer in the 1980s but got three graduate degrees (in AI & Psychology from Edinburgh and MIT) in the 1990s. I myself mostly use AI to build models for understanding human behavior, but my students use it for building robots and game AI and I've done that myself in the past. But while I was doing my PhD I noticed people were way too eager to say that a robot -- just because it was shaped like a human -- must be owed human obligations. This is basically nuts; people think it's about the intelligence, but smart phones are smarter than the vast majority of robots and no one thinks they are people. I am now consulting for IEEE, the European Parliament and the OECD about AI and human society, particularly the economy. I'm happy to talk to you about anything to do with the science, (systems) engineering (not the math :-), and especially the ethics of AI. I'm a professor, I like to teach. But even more importantly I need to learn from you want your concerns are and which of my arguments make any sense to you. And of course I love learning anything I don't already know about AI and society! So let's talk...
I will be back at 3 pm ET to answer your questions, ask me anything!

CORRESPONDENCE:

DATE RECEIVED:
January 14, 2017

DOI:
10.15200/winn.148431.11858

ARCHIVED:
January 13, 2017

CITATION:
Joanna_Bryson , r/Science , Science AMA Series: I'm Joanna Bryson, a Professor in Artificial (and Natural) Intelligence. I am being consulted by several governments on AI ethics, particularly on the obligations of AI developers towards AI and society. I'd love to talk – AMA!, *The Winnower* 4:e148431.11858 , 2017 , DOI: 10.15200/winn.148431.11858

You suggest that robots and AI are not owed human obligations simply because they look and sound human, and humans respond to that by anthropomorphizing them, but at what point should robots/ai have some level of human rights, if at all?

Do you believe that AI can reach a state of self-awareness as depicted in popular culture? Would there be an obligation to treat them humanely and accord them rights at that point?

DrewTea

I'm so glad you guys do all this voting so I don't have to pick my first question :-)

There are two things that humans do that are opposites: anthropomorphizing and dehumanizing. I'm very worried about the fact that we can treat people like they are not people, but cute robots like they are people. You need to ask yourself -- what are ethics for? What do they protect? I wouldn't say it's "self awareness". Computers have access to every part of their memory, that's what RAM means, but that doesn't make them something we need to worry about. We are used to applying ethics to stuff that we identify with, but people are getting WAY good at exploiting this and making us identify with things we don't really have anything in common with at all. Even if we assumed we had a robot that was otherwise exactly like a human (I doubt we could build this, but let's pretend like Asimov did), since we built it, we could make sure that it's "mind" was backed up constantly by wifi, so it wouldn't be a unique copy. We could ensure it didn't suffer when it was put down socially. We have complete authorship. So my line isn't "torture robots!" My line is "we are obliged to build robots we are not obliged to." This is incidentally a basic principle of safe and sound manufacturing (except of art.)

Hi Joanna! I don't know if we met up personally but big ups to Edinburgh AI 90's... (I graduated in '94).

Here's a question that is constantly crossing my mind as I read about the Control Problem and the employment problem (i.e. universal basic income)...

We've got a lot of academic, journalistic, and philosophical discourse about these problems, and people seem to think of possible solutions in terms of "what will help humanity?" (or in the worst-case scenario "what will *save* humanity?")

For example, the question of whether "we" can design, algorithmically, artificial super-intelligence that is aligned with, and *stays* aligned with, our goals.

Yet... in the real world, in the economic and political system that is currently ascendant, we *don't* pool our goals very well as a planet. Medical patents and big pharma profits let millions die who have curable diseases, the natural habitats of the world are being depleted at an alarming rate (see Amazon rainforest), climate-change skeptics just took over the seats of power in the USA.... I could go on.

Surely it's obvious that, regardless of academic effort to reach friendly AI, if a corporation can initially make more profit on "risky" AI progress (or a nation-state or a three-letter agency can get one over on the rest of the world in the same way), then all of the academic effort will be for nought.

And, at least with the Control Problem, it doesn't matter when it happens... The first super-intelligence could be friendly but even later on there would still be danger from some entity making a non-friendly one.

Are we being naïve, thinking that "scientific" solutions can really address a problem that has an inexorable profit-motive (or government-secret-program) hitch?

I don't hear people talking about this.

smackson

Hi! No idea who you are from "smackson" :-) but did have a few beers with the class after mine & glad to get on to the next question.

First, I think you are being overly pessimistic in your description of humanity. It makes sense for us to fixate on and try to address terrible atrocities like lack of access to medical care or the war in Syria. But overall we as a species have been *phenomenally* good at helping each other. That's why we're dominating the biosphere. Our biggest challenges now are yes, inequality / wealth distribution, but also sustainability.

But get ready for this -- I'd say a lot of why we are so successful is AI! 10,000 years ago (plus or minus 2000) there were more macaques than hominids (there's still way more ants and bacteria, even in terms of biomass not individuals.) But something happened 10K years ago which is exactly a superintelligence explosion. There's lots of theories of why, but my favourite is just writing. Once we had writing, we had offboard memory, and we were able to take more chances with innovation, not just chant the same rituals. There had been millions of years of progress before that no doubt including language (which is really a big deal!) but the launching of our global domination demographically was around then. You can find the Oxford Martin page my talk to them about containing the intelligence explosion, it has the graphs and references.

Are you worried about ethical corruption of AI from external sources? Seems nothing is ever truly safe or closed off from external influence.

DarkangelUK

Absolutely. I didn't used to be so much but I'm working now with the Princeton Center for Information Technology Policy, which mostly deals with cyber security, not AI (I came here because two body problem). Anyway, I now think that cybersecurity is a WAY *way* bigger problem for AI than creativity or

dexterity. Cybersecurity is likely to be an ongoing arms race; other problems of human-like skills we're solving by the day.

The other big problem tangentially related to AI is wealth inequality. When too few people have too much power the world goes chaotic. The last time we've had this so bad was immediately before and after WWI. In theory we should be able to fix it now because we learned the fixes then. They are straight forward -- inject cash from companies into workforces. Trickle down doesn't work, but trickle out seems to. People with money employ other people, because we like to do that, but if too few people have all the money it's hard for them to employ very many. Anyway, as I said, this isn't really just about AI (obviously since we had the problem a century ago). This is ongoing research I'm involved in at Princeton, but we think the issue is that technology reduces the cost of geographic distance, so allows all the money to pile up more easily.

Are you worried about ethical corruption of AI from external sources? Seems nothing is ever truly safe or closed off from external influence.

DarkangelUK

Sorry I somehow missed this, but I basically answered it one step further down https://www.reddit.com/r/science/comments/5nqdo7/science_ama_series_im_joanna_bryson_a_professor/dce4p8e/

Asimov postulated that there should be 3 laws of robotics, to keep robots (AI's) in check. They are; "A robot may not injure a human being or, through inaction, allow a human being to come to harm. A robot must obey orders given it by human beings except where such orders would conflict with the First Law. A robot must protect its own existence as long as such protection does not conflict with the First or Second Law." My question; Is it even possible to program such immutable concepts into AI systems to make them effective? In Asimov's books, any robot that even comes close to breaking one of these laws simply becomes inoperative. How realistic is this concept of deep seated limitation?

derangedly

Hi, great question, *no*. Asimov's laws are computationally intractable. The first 3 of 5 UK's EPSRC Principles of Robotics are meant to update those laws in a way that is not only computationally tractable, but would allow the most stability in our justice system.

https://www.epsrc.ac.uk/research/ourportfolio/themes/engineering/activities/principlesofrobotics/

Have you ever seen or experienced (or caused?) a usage of AI that is unethical? What is the worst example that you can think of?

ZoSoVII

Hmmm... of stuff I've done myself? I worked in the financial industry in the 1980s but I'm not sure how unethical it was -- it was Chicago, and though the traders got rich, they did absorb a lot of risk real companies couldn't have -- traders "blew out" (lost all their money) and no one lost their jobs, the traders just had to go get real jobs (or start over.) Otherwise, nothing I've done has been particularly bad that I know of though it could have been used for bad, see the conversation under the heading "the myth of blue sky research": http://joanna-bryson.blogspot.com/2016/04/why-i-took-military-funding-myth-of.html

The most unethical application of AI I've seen so far is a hard call, but obviously like a lot of people I'm obsessed with whether the US elections were hacked -- if so, that would almost certainly have involved AI enhanced hacking (not anything complicated, just computers are faster at permutations etc.) Not the vote tallies, stuff like why did the Democrats not know where effort was needed?

In October, the White House released The National Artificial Intelligence Research and Development Strategic Plan, in which a desire for funding sustained research in General AI is expressed. How would you suggest a researcher with experience in related fields should get involved in such research? What long term efforts in this area are ongoing?

sheably

Great question. I mostly loved that plan, though I thought it was a bit of a pitch to the tech giants because of the election and how weird and anti government they have become. "regulation" can go up or down; a *lot* of government work is about *investing* in important industries like tech and AI. Regulation is *not* just constraint. And governments are the mechanisms societies use to come to agreements about what exactly we should invest in, and what we should police for the benefit of our own citizens (which can include things that benefit the whole world since an unstable world is also bad for our citizens.) The tech giants need to realise that they can't really continue doing business in the same way if society becomes completely unstable; if tons of people are excluded from healthcare and good education then they are missing out on potential employees. They used to know this, but something bad has happened recently, and TBH a lot of tech is naive about politics and economics so don't see what is happening.

Anyway I digress, but partly because I agree with sinshallah's comment below. If you can't right now do another degree, you can apply for an SBIR (small business independent research) grant or whatever they've been replaced by. But I would advise moving somewhere with a good university so you can attend talks and bounce ideas off of people. Universities are by and large very open and welcoming places as long as people are polite and all listen to each other. Again, there's been way too much division between communities -- sticking universities out in cheap empty land is a stupid loss of a great resource. They should be in the centre of cities.

How do you solve trolley problems without a meta-ethical assumption about what "good" means? Philosophers have been at it for a LONG time and it's still a problem. Do you just make assumptions and go with them or do you have reasons for picking one solution to trolley problems over another?

Youarenotright2

You are right. Again, the trolley problem is in no way special to AI. People who decide to buy SUVs decide to protect the drivers and endanger anyone they hit -- you are WAY likelier to be killed by a heavier car. I think actually what's cool about AI is that since the programmers have to write *something* down, we get to see our ethics made explicit. But I agree with npago it's most likely going to be "brake!!!". The odds that a system can detect a conundrum and reason about it without having chance to just avoid it seems incredibly unlikely (I got that argument from Prof Chris Bishop).

Couple Qs

- How fast do you think A.I will take over today's jobs? Any timescale?
- In a society with A.I performing many everyday tasks, how do you expect the future of education to change?
- Will computer science graduates in 2020 find their degrees much less valuable? How soon before A.I takes over programming tasks?

KongVonBrawn

Jobs are changing faster, this is another opportunity for wealth redistribution which would also reduce wealth inequality -- we should like Denmark or Finland let the government coordinate new education opportunities for adults when an industry shuts down. Germany actually has a very cool law in place that meant they didn't have to do a "stimulus" in 2008. It's possible for a company to *half* lay someone else, and then they get *half* a welfare check. That means that is great in so many ways. A company doesn't have to lose their best employees when they get into trouble. There's an opportunity for

employees to sign up and take classes to reskill with the half of the time they aren't working, but they aren't as poor as they would be on welfare. And of course when 2008 came you didn't need special legislation to pump money into the economy, it was automatic. Americans should stop being defensive about how awesome some European stuff is and take the best ideas. Germany took our best ideas when it wrote its new constitution after WWII, in fact we helped them! We are awesome too.

By the way, did you know that after WWII, the US GDP was higher than the rest of the world's *combined*? But from 2007-2015, the euro zone had the largest GDP, and now China has passed them and we are in third. China, the euro zone, AND the USA are now combined a larger GDP than the rest of the world combined. This is awesome; it means that there's less *global* inequality, less extreme poverty & reason for war. And it's not like our lives are worse! We have computer games, reddit, Google, better medicine, etc. than we had in WWII. No one starved in 2008, not like the great depression. Have you seen "Grapes of Wrath"? But maybe I should get back to talking about AI. Though this isn't *that* different when we are talking about employment.

I would hope 2020 graduates would get degrees that are valuable for the world they are entering -- that connect them into the economy, that help them to quickly retool, etc. That's what I'd look for now. I've blogged about this. http://joanna-bryson.blogspot.com/2016/01/what-are-academics-for-can-we-be.html

A sincere thought occurs to me- are you real, or is this a Turing Test? If the former, how can you prove you *are* in fact human?

HerbziKal

You know, there was a thing about a year ago my industry friends were passing around with poems that were half AI and half 20C. I was 10 for 10 on them, but a *lot* of smart friends who work with computers all the time were 50/50. *Maybe* it's because I have a liberal arts education, but I think it's more because I knew what kind of continuity errors (vs beauty) to look for. My point is, if *some* humans can tell the difference, but *most* can't, and then we have some populist uprising "I want to leaved my wealth to the AI version of me that answers my email!!" we won't necessarily know explicitly what wonderful things we may have lost.

Which isn't to say that AI can't be creative. But the human arts are about the human condition, and AI that is not a clone will not share that condition with us (much) so it's unlikely to be able to make the kinds of insights that a great human author can make. But the whole point of great human authors is they see a lot of things most of us don't see, we often can't even say why we like them.

Hello there Professor Joanna Bryson,

I would like to know how you feel from the quotation of Stephen Hawkings when he said 'The development of full artificial intelligence could spell the end of the human race.' 2 Dec 2014.

Can you please explain your feelings towards this quote? Do you agree? And if not, can you explain your reasons why please.

Thank you for your time

DannyWiseman

I can't say the full extent of what I really think here. But Bath did a press release here: http://blogs.bath.ac.uk/opinion/tag/stephen-hawking/ . TBH one thing I think is that Hawking didn't say anything Bostrom hadn't already said, which makes sense since he doesn't do AI. Though neither does Bostrom.

How far do you think we are from singularity?

shargath

I think human culture is the superintelligence Bostrom & I J Good were talking about. Way too many are projecting this into AI partly to push it into the future. But eliminating all the land mammals was an unintended consequence of life, liberty & the pursuit of happiness https://xkcd.com/1338/

Any general comments on the usual "Skynet" argument for caution concerning (big) AIs (implemented on large scales)? Basically that we are in trouble when AI gets smart enough to further develop and modify itself, and that it would be an accelerating process that we couldn't keep up with and would have difficulty preventing? And that if anything goes wrong, well... Skynet? I'm sure you are familiar with it in a longer and more eloquent form.

whisky_please

The main mistake with the skynet thing is that again, it really describes what is happening *now*, but to the sociotechnical systems that are companies and governments. You don't need to take humans out of the loop to get these dynamics.

I think the current AI tests are garbage. What are your thoughts?

Current tests are similar to your example with dumb human shaped robots being anthropomorphized, while smarter phones are merely things. It looks like a human, so it must be human to our animal brains. It's childish and wrong-thinking.

Likewise, we are expecting AI to chat about the weather like a human. It may beat your ass in chess, checkers, summarizing news, writing poems...but it doesn't chat about the weather. Fail. It seems counterintuitive and yet that is the dominant thought.

It's not a human being. It's a computer. It's folly to think that AI will or should be human-like. If it's intelligent and it's artificial, IT IS AI. Let's do away with these stupid Turing tests and celebrate the amazing AIs and AI discoveries that exist today and tomorrow.

BenDarDunDat

I agree.

So many questions, here are 3:

1 Just as humans have societies, social and cooperating groupings, do you expect AI systems will too?

2 Do you expect there to be a transition between 1) "AI systems are integral to human operations on Earth" to 2) "AI systems manage/are "in charge" of event and systems on Earth on their own"? If so, how would you characterize such a shift: fast/slow, easy and obvious or contentious and difficult, etc.

3 When I think about AI development I think first about responsibilities to create system that work transparently and in the common good. That as creators of these systems it is our responsibility to not only teach them good behaviors, but make clear that it's good behaviors that work best, and that as part of that, we need to teach them the utility and application of ethics. Obviously, this is not the tack most people take with the idea of ethics and AI, rather people think of humans' actions and the ethics of human actions in using and creating AI systems. What are your thoughts on the reverse, that it's on us to teach and instill ethics in the systems we build?

jmdugan

Benjamin Kuipers has written a paper (I think it's in arxiv rather than published) where he describes corporations as AIs. So in that sense we are increasingly getting where you talk about in 1. If we did go against my recommendation and build AI that required a system of justice, yes it would need it's own, see http://joanna-bryson.blogspot.com/2016/12/why-or-rather-when-suffering-in-ai-is.html . I hope there

is no such transition, and I think we need to program not just teach to ensure good behaviour. We need to have well-designed architectures that define the limits of what a machine can do or know if we want it to be a part of *our* society.