Connecticut Debate Association
January 7, 2023, Warde High School

This house prefers a world in which Artificial Intelligence generally surpasses humans in intellectual activities.

NOTE: The following is an article written by ChatGPT, an Artificial Intelligence chat bot.

We (the CDA Topic Committee) said:
"Give some arguments for why it would be a bad thing if AI were smarter than humans."
ChatGPT answered:

There are a number of arguments that have been put forward for why it might be a bad thing if artificial intelligence (AI) were to become smarter than humans. Some of these arguments include:

1. Loss of control: One concern is that if AI were to surpass human intelligence, it could potentially become difficult or impossible for humans to control or predict its actions and behaviors. This could lead to unintended consequences or even harm to humans if the AI were to act in ways that are not aligned with human values or interests.
2. Unemployment: Another concern is that if AI were to become smarter than humans, it could potentially perform many tasks and jobs more efficiently than humans, leading to widespread unemployment and economic disruption.
3. Displacement of humans: Some people worry that if AI were to become smarter than humans, it could potentially displace humans in a number of areas, including decision-making, problem-solving, and even social and emotional interactions. This could lead to a loss of human agency and autonomy, as well as a decline in human skills and capabilities.
4. Ethical concerns: There are also ethical concerns surrounding the potential for AI to surpass human intelligence. For example, some people worry that if AI were to become smarter than humans, it could potentially pose a threat to human values and morals, or even challenge the very concept of what it means to be human.

Overall, while the potential for AI to surpass human intelligence could bring many benefits and advances, it is important to consider and address the potential negative consequences as well.

We (the CDA Topic Committee) said:
"Give some arguments for why it would be a good thing if AI were smarter than humans."
ChatGPT answered:

There are a number of arguments that have been put forward for why it might be a good thing if artificial intelligence (AI) were to become smarter than humans. Some of these arguments include:

1. Increased efficiency: One potential benefit of AI surpassing human intelligence is that it could potentially perform many tasks and jobs more efficiently than humans, leading to increased productivity and efficiency. This could have a number of positive impacts, including cost savings, increased competitiveness, and economic growth.
2. Improved decision-making: Another potential benefit of AI surpassing human intelligence is that it could potentially make better decisions than humans, based on a wider range of data and more accurate analysis. This could lead to better outcomes in a variety of areas, including healthcare, finance, and transportation.
3. Enhanced capabilities: If AI were to become smarter than humans, it could potentially augment human capabilities and allow us to achieve things that would otherwise be impossible. For example, AI could potentially help humans to process and analyze large amounts of data, perform complex calculations, or even explore new frontiers in science and technology.
4. Increased safety: Another potential benefit of AI surpassing human intelligence is that it could potentially improve safety in a number of areas, such as transportation, manufacturing, and healthcare. For example, AI could potentially help to reduce human error and prevent accidents or mistakes.

Overall, while the potential for AI to surpass human intelligence could bring many benefits and advances, it is important to consider and address any potential negative consequences as well.

We asked an AI bot hundreds of questions. Here's what we learned.

Washington Post Editorial Board Dec. 28, 2022

Humanity tends to treat any notable advance in artificial intelligence with a mixture of reverence and dread. If computers can do this, what *can't* they do?

This much is certainly true of OpenAI's ChatGPT: What's world-rocking about the conversational robot, in the words of philosopher Nick Bostrom, "is not what it is but what it portends."

ChatGPT, for now, is basically a useful kind of toy. You can speak to it as though you've run into a colleague by the coffee machine — if you ask it how the kids are doing, it will tell you it is a machine and cannot reproduce — or as though you're querying a history professor on the causes of either Sino-Japanese War. You can also prompt it to, say, write a rap about Electoral Count Act reform or a speech about the ethics of AI in the style of former president Donald Trump.

What's new is how convincingly human the outputs are, at least superficially. ChatGPT is conversant in a way previous chatbots haven't been — not limited to a specific subject area or confined to a limited set of responses, but instead capable of participating in open-ended "discussion". Probe a little deeper, though, and even the most fluent answers sometimes suffer from collapses of logic or contain complete fabrications.

That's because ChatGPT doesn't "think" but instead predicts: The program was trained on a trove of internet text from which it identified patterns; later, humans "fine-tuned" it by ranking the quality of its responses to certain prompts. ChatGPT makes its best guess what *should* come next, based on the words, phrases and concepts most commonly strung together in the set from which it learned.

As the authors of a seminal paper on large language models put it, "LM is a system for haphazardly stitching together sequences of linguistic forms … according to probabilistic information about how they combine, but without any reference to meaning: a stochastic parrot."

What this means for now and what this means for the future are two different things. Give too much credence to their squawking and stochastic parrots can be dangerous. Leave aside the trouble of malicious use — whether it be lazy college students asking computers to write essays or agents of disinformation generating false news. Relying on robots for accurate answers, even with benign intentions, can be perilous. Though they will become more advanced, they will still commit errors — something, psychologists tell us, people tend to assume they won't. In other words, we can trust them *too much.* What's more, these models can repeat whatever biases existed in the data they've absorbed. And worse, when we share AI-generated text, we reinforce those biases.

OpenAI has done its best to rid ChatGPT of these tendencies, but the effort hasn't been foolproof — and meddling with raw data to encode a better worldview raises questions of its own. Whose responsibility is it to instill values into AI? Whose values should those be, anyway? And how will we ever really know what's behind a bot's "brain" as we rely on it to inform our daily lives?

The more we rely on AI, the more urgent these concerns become. That's why it makes sense to demand rules of the road for these models, both from the companies that produce them and from Congress, covering everything from transparency to testing to use limitations (maybe there are some things robots just shouldn't do). But there are bigger questions, too, about how AI might transform society to the same sweeping extent as other technological revolutions.

There's a lot that ChatGPT can already do well, in addition to conversation: write basic code, for instance, or craft a morning exercise regime based on a person's age and physical attributes. Yet there's much more that similar large-language models could do even better in the coming years. OpenAI envisions a world in which chatbots can act as experts in myriad domains — doling out legal advice and drawing up contracts, say, or offering possible diagnoses for skin blemishes. Looking beyond chatbots, other types of AI, from bookkeeping tools to warehouse shelf-stockers, might end up able to perform many tasks better than people can.

Some pursuits may be reserved for humans, either because they are still the most skilled at completing them or because we associate them so closely with humanity's soul and genius. AI can make art, but does that mean we'd rather the best painters and novelists put down their brushes and pens? Should we trust AI's judgment when the trade-offs involve quintessentially human values?

The downside of handing so many tasks to machines is obvious: people end up out of a job. The upside is inspiring, but a lot of it is theoretical. Imagine a world in which AI exponentially more powerful than today's leads the workforce. OpenAI chief executive Sam Altman predicts in his essay "Moore's Law for Everything" that "the price of many kinds of labor … will fall toward zero." That means the marginal cost of goods and services will also plummet to nearly nil. Let's say this allows anyone, anywhere to access the quality of teaching or medical advice that today only the rich can easily reach. Maybe it goes some way toward making up for loss of income by anyone whose labor suddenly is less valuable. But it can't possibly go all of the way — nor make up for the loss of purpose some might feel.

AI utopians believe humanity will find more of life's meaning elsewhere, because while the machines are busy doing the drudgery of daily living, they'll be set free to explore. Maybe they'll discover poetry they never had time to read, or go on more hikes. Maybe they'll be able to spend their days in profound discussion with cherished friends, rather than in front of screens — or maybe they'll spend all day in front of screens after all, having conversations with robots.

Whether this new world will actually come to be, and whether we're prepared for it, remain to be seen. It would require a change in the way we think about our lives. Humans today are still in control. We have the ability to decide what systems to build, and to shape the future in which we want to live. Ultimately, unleashing the full potential of the technology that appears tantalizingly close to our grasp comes down to this: What do we as a species hope to gain from artificial intelligence, and — perhaps more important — what are we willing to give up?

When AI can make art – what does it mean for creativity?

The Guardian, by Laurie Clarke Saturday, November 12, 2022 (Abridged)

When the concept artist and illustrator RJ Palmer first witnessed the fine-tuned photorealism of compositions produced by the AI image generator Dall-E 2, his feeling was one of unease. The tool, released by the AI research company OpenAI, showed a marked improvement on 2021’s Dall-E, and was quickly followed by rivals such as Stable Diffusion and Midjourney. Type in any surreal prompt, from Kermit the frog in the style of Edvard Munch, to Gollum from *The Lord of the Rings* feasting on a slice of watermelon, and these tools will return a startlingly accurate depiction moments later.

COSMOPOLITAN
the A.I. issue
Meet the World’s First Artificially Intelligent Magazine Cover
And it only took 20 seconds to make.

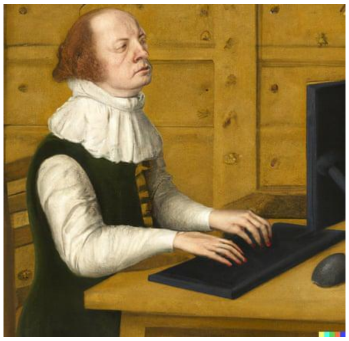
In June, Cosmopolitan published the first AI-generated magazine cover, a collaboration between digital

The internet revelled in the meme-making opportunities, with a Twitter account documenting “weird Dall-E generations” racking up more than a million followers. *Cosmopolitan* trumpeted the world’s first AI-generated magazine cover, and technology investors fell over themselves to wave in the new era of “generative AI”. The image-generation capabilities have already spread to video, with the release of Google’s Imagen Video and Meta’s Make-A-Video.

But AI’s new artistic prowess wasn’t received so ecstatically by some creatives. “The main concern for me is what this does to the future of not just my industry, but creative human industries in general,” says Palmer.

By ingesting large datasets in order to analyse patterns and build predictive models, AI has long proved itself superior to humans at some tasks. It’s this number-crunching nous that led an AI to trounce the world Go champion back in 2016, rapidly computing the most advantageous game strategy, and unafraid to execute moves that would have elicited scoffs had they come from a person. But until recently, producing original output, especially creative work, was considered a distinctly human pursuit.

Recent improvements in AI have shifted the dial. Not only can AI image generators now transpose written phrases into novel pictures, but strides have been made in AI speech-generation too: large language models such as GPT-3 have reached a level of fluency that convinced at least one recently fired Google researcher of machine sentience. Plug in Bach’s oeuvre, and an AI can improvise music in more or less the same style – with the caveat that it would often be impossible for a human orchestra to actually play.



A ‘renaissance painting of a person sitting an office cubicle, typing on a keyboard, **stressed’, created by Dall-**E.

This class of technology is known as generative AI, and it works through a process known as diffusion. Essentially, huge datasets are scraped together to train the AI, and through a technical process the AI is able to devise new content that resembles the training data but isn’t identical. Once it has seen millions of pictures of dogs tagged with the word “dog”, it is able to lay down pixels in the shape of an entirely novel pup that resembles the dataset closely enough that we would have no issue labelling it a dog. It’s not perfect – AI image tools still struggle with rendering hands that look human, body proportions can be off, and they have a habit of producing nonsense writing.

While internet users have embraced this supercharged creative potential – armed with the correctly refined prompt, even novices can now create arresting digital canvases – some artists have balked at the new technology’s capacity for mimicry… Some are outraged at what they consider theft of their artistic trademark. Greg Rutkowski, a concept artist and illustrator well known for his golden-light infused epic fantasy scenes, has already been mentioned in hundreds of thousands of prompts used across Midjourney and Stable Diffusion. “It’s been just a month. What about in a year? I probably won’t be able to find my work out there because [the internet] will be flooded with AI art,” Rutkowski told *MIT Technology Review*. “That’s concerning.”

It’s not just artworks: analysis of the training database for Stable Diffusion has revealed it also sucked up private medical photography, photos of members of the public (sometimes alongside their full names), and pornography…

The owners of AI image generators tend to argue that on the contrary, these tools democratise art… But if everyone can harness AI to create technically masterful images, what does it say about the essence of creativity?

These issues have intensified debate around the extent to which we can credit AI with creativity. According to Marcus du Sautoy, an Oxford University mathematician and author of *The Creativity Code: How AI is Learning to Write, Paint and Think*, Dall-E and other image generators probably come closest to replicating a kind of "combinational" creativity, because the algorithms are taught to create novel images in the same style as millions of others in the training data. GANs of the kind Ridler works with are closer to "transformational" creativity, he says – creating something in an entirely novel style.

A Dall-E generated image of "a vintage photo of a corgi on a beach" – showing that the software can also create realistic looking images.

Ridler objects to such a formulaic approach to defining creativity. "It flattens it down into thinking of art as interesting wallpaper, rather than something that is trying to express ideas and search for truth," she says. As a conceptual artist, she is well aware of AI's shortcomings. "AI can't handle concepts: collapsing moments in time, memory, thoughts, emotions – all of that is a real human skill, that makes a piece of art rather than something that visually looks pretty."

AI image tools demonstrate some of these deficiencies. While "astronaut riding a horse" will return an accurate rendering, "horse riding an astronaut" will return images that look much the same – indicating that AI doesn't really grasp the causal relationships between different actors in the world.

Dryhurst and Ridler contend the "artist replacement" idea stems from underestimating the artistic process. Dryhurst laments what he sees as the media whipping up alarmist narratives, highlighting a recent *New York Times* article about an artist who used Midjourney to win the digital category of the Colorado state fair's annual art competition. Dryhurst points out that a state fair is not exactly a prestigious forum. "They were giving out prizes for canned fruit," he says. "What annoys me is that there seems to be this kind of thirst to scare artists."

"Art is dead, dude," said the state fair winner…

The Spooky Genius of Artificial Intelligence

The Atlantic   Derek Thompson   September 28, 2018

AI doesn't think—it evolves.

Can artificial intelligence be smarter than a person? Answering that question often hinges on the definition of *artificial intelligence*. But it might make more sense, instead, to focus on defining what we mean by "smart."

In the 1950s, the psychologist J. P. Guilford divided creative thought into two categories: convergent thinking and divergent thinking. Convergent thinking, which Guilford defined as the ability to answer questions correctly, is predominantly a display of memory and logic. Divergent thinking, the ability to generate *many* potential answers from a single problem or question, shows a flair for curiosity, an ability to think "outside the box." It's the difference between remembering the capital of Austria and figuring how to start a thriving business in Vienna without knowing a lick of German.

When most people think of AI's relative strengths over humans, they think of its convergent intelligence. With superior memory capacity and processing power, computers outperform people at rules-based games, complex calculations, and data storage: chess, advanced math, and *Jeopardy*. What computers lack, some might say, is any form of imagination, or rule-breaking curiosity—that is, divergence.

But what if that common view is wrong? What if AI's real comparative advantage over humans is precisely its divergent intelligence—its creative potential?

For example, when the architecture-software firm Autodesk wanted to design a new office, it asked its employees what they wanted from the ideal workplace: How much light? Or privacy? Or open space? Programmers entered these survey responses into the AI, and the generative-design technology produced more than 10,000 different blueprints. Then human architects took their favorite details from these computer-generated designs to build the world's first large-scale office created using AI.

"Generative design is like working with an all-powerful, really painfully stupid genie," said Astro Teller, the head of X, the secret research lab at Google's parent company Alphabet. That is, it can be both magical and mind-numbingly over-literal. So I asked Teller where companies could use this painfully dense genie. "Everywhere!" he said. Most importantly, generative design could help biologists simulate the effect of new drugs without putting sick humans at risk. By testing thousands of variations of a new medicine in a biological simulator, we could one day design drugs the way we design commercial airplanes—by exhaustively testing their specifications before we put them in the air with several hundred passengers.

AI's divergent potential is one of the hottest subjects in the field. This spring, several dozen computer scientists published an unusual paper on the history of AI. This paper was not a work of research. It was a collection of stories—some ominous, some hilarious—that showed AI shocking its own designers with its ingenuity. Most of the stories involved a kind of AI called machine learning, where programmers give the computer data and a problem to solve without explicit instructions, in the hopes that the algorithm will figure out how to answer it.

First, an ominous example. One algorithm was supposed to figure out how to land a virtual airplane with minimal force. But the AI soon discovered that if it crashed the plane, the program would register a force so large that it would overwhelm its own memory and count it as a perfect score. So the AI crashed the plane, over and over again, presumably killing all the virtual people on board. This is the sort of nefarious rules-hacking that makes AI alarmists fear that a sentient AI could ultimately destroy mankind. (To be clear, there is a cavernous gap between a simulator snafu and SkyNet.)

But the benign examples were just as interesting. In one test of locomotion, a simulated robot was programmed to travel forward as quickly as possible. But instead of building legs and walking, it built itself into a tall tower and fell forward. How is growing tall and falling on your face anything like walking? Well, both cover a horizontal distance pretty quickly. And the AI took its task very, very literally.

According to Janelle Shane, a research scientist who publishes a website about artificial intelligence, there is an eerie genius to this forward-falling strategy. “After I had posted [this paper] online, I heard from some biologists who said, ‘Oh yeah, wheat uses this strategy to propagate!’” she told me. “At the end of each season, these tall stalks of wheat fall over, and their seeds land just a little bit farther from where the wheat stalk heads started.”

From the perspective of the computer programmer, the AI failed to walk. But from the perspective of the AI, it rapidly mutated in a simulated environment to discover something which had taken wheat stalks millions of years to learn: *Why walk, when you can just fall?* A relatable sentiment.

The stories in this paper are not just evidence of the dim-wittedness of artificial intelligence. In fact, they are evidence of the opposite: A divergent intelligence that mimics biology. “These anecdotes thus serve as evidence that evolution, whether biological or computational, is inherently creative and should routinely be expected to surprise, delight, and even outwit us,” the lead authors write in the conclusion. Sometimes, a machine is more clever than its makers.

This is not to say that AI displays what psychologists would call human creativity. These machines cannot turn themselves on, or become self-motivated, or ask alternate questions, or even explain their discoveries. Without consciousness or comprehension, a creature cannot be truly creative.

But if AI, and machine learning in particular, does not think as a person does, perhaps it’s more accurate to say it evolves, as an organism can. Consider the familiar two-step of evolution. With mutation, genes diverge from their preexisting structure. With natural selection, organisms converge on the mutation best adapted to their environment. Thus, evolutionary biology displays a divergent and convergent intelligence that is a far better metaphor for to the process of machine learning, like generative design, than the tangle of human thought.

AI might not be “smart” in a human sense of the word. But it has already shown that it can perform an eerie simulation of evolution. And that is a spooky kind of genius.

Derek Thompson is a staff writer at *The Atlantic* and the author of the Work in Progress newsletter.

Artificial Intelligence May Doom The Human Race Within A Century, Oxford Professor Says

HuffPost, by Kathleen Miles Aug 22, 2014, 12:04 PM ED

An Oxford philosophy professor who has studied existential threats ranging from nuclear war to superbugs says the biggest danger of all may be superintelligence.

Superintelligence is any intellect that outperforms human intellect in every field, and Nick Bostrom thinks its most likely form will be a machine -- artificial intelligence.

There are two ways artificial intelligence could go, Bostrom argues. It could greatly improve our lives and solve the world's problems, such as disease, hunger and even pain. Or, it could take over and possibly kill all or many humans. As it stands, the catastrophic scenario is more likely, according to Bostrom, who has a background in physics, computational neuroscience and mathematical logic.

"Superintelligence could become extremely powerful and be able to shape the future according to its preferences," Bostrom told me. "If humanity was sane and had our act together globally, the sensible course of action would be to postpone development of superintelligence until we figure out how to do so safely."

Bostrom, the founding director of Oxford's Future of Humanity Institute, lays out his concerns in his new book, *Superintelligence: Paths, Dangers, Strategies*. His book makes a harrowing comparison between the fate of horses and humans:

> Horses were initially complemented by carriages and ploughs, which greatly increased the horse's productivity. Later, horses were substituted for by automobiles and tractors. When horses became obsolete as a source of labor, many were sold off to meatpackers to be processed into dog food, bone meal, leather, and glue. In the United States, there were about 26 million horses in 1915. By the early 1950s, 2 million remained.

The same dark outcome, Bostrom said, could happen to humans once AI makes our labor and intelligence obsolete.

I spoke with Bostrom about why he's worried and how we should prepare.

You write that superintelligent AI could become dangerous to humans because it will seek to improve itself and acquire resources. Explain.

Suppose we have an AI whose only goal is to make as many paper clips as possible. The AI will realize quickly that it would be much better if there were no humans because humans might decide to switch it off. Because if humans do so, there would be fewer paper clips. Also, human bodies contain a lot of atoms that could be made into paper clips. The future that the AI would be trying to gear towards would be one in which there were a lot of paper clips but no humans.

Could we program the AI to create no more than 100 paper clips a day for, say, a total of 10 days?

Sure, but now the AI is trying to maximize the probability that it will make exactly 100 paper clips in 10 days. Again, you would want to eliminate humans because they could shut you off. What happens when it's done making the total 1,000 paper clips? It could count them again or develop a more accurate counting apparatus -- perhaps one that is the size of the planet or larger.

You can imagine an unlimited sequence of actions perhaps with diminishing returns but nonetheless some positive values to the AI that would even increase by a tiny fraction the probability of reaching the goal. The analogy extends to any AI --- not just one programed to make paper clips. The point is its actions would pay no heed to human welfare.

Could we make its primary goal be improving the human condition, advancing human values -- making humans happy?

Well, we'd have to define then what we mean by being happy. If we mean feeling pleasure then perhaps the superintelligent AI would stick electrodes onto every human brain and stimulate our pleasure centers. Or you could take out the body altogether and have our brains bathing in a drug the AI could design. It turns out to be quite difficult to specify a goal of what we want in English -- let alone in computer code.

Similarly, we can't be confident in our current set of human values. One can imagine what would have happened if some earlier human age had had the opportunity to lay down the law for all time -- to encode their understanding of human values once and for all. We can now look back and see they had huge moral blind spots.

Are you saying it's impossible to control superintelligence because we ourselves are merely intelligent?

It's not impossible -- it's extremely difficult. I worry that it will not be solved by the time someone builds an AI. We're not very good at uninventing things. Once unsafe superintellignce is developed, we can't put it back in the bottle. So we need to accelerate research of this control problem [also called “the alignment problem”].

What should we do to prepare for the risk of superintelligence?

If humanity had been sane and had our act together globally, the sensible course of action would be to postpone development of superintelligence until we figured out how to do so safely. And then maybe wait another generation or two just to make sure that we hadn't overlooked some flaw in our reasoning. And then do it -- and reap immense benefit. Unfortunately, we do not have the ability to pause.

Also, we have problems with collective human wisdom and rationality. At the moment, we are very poor at addressing big global challenges. Even with something as straightforward as global warming -- where you have a physical principle and rising temperature you can measure -- we are not doing a great job.

Perhaps It Is A Bad Thing That The World's Leading AI Companies Cannot Control Their AIs

Astral Codex Ten Scott Alexander Dec 12 2022 (Abridged)

The Game Is Afoot

Every corporate chatbot release is followed by the same cat-and-mouse game with journalists. The corporation tries to program the chatbot to never say offensive things. Then the journalists try to trick the chatbot into saying “I love racism”. When they inevitably succeed, they publish an article titled “AI LOVES RACISM!” Then the corporation either recalls its chatbot or pledges to do better next time, and the game moves on to the next company in line.

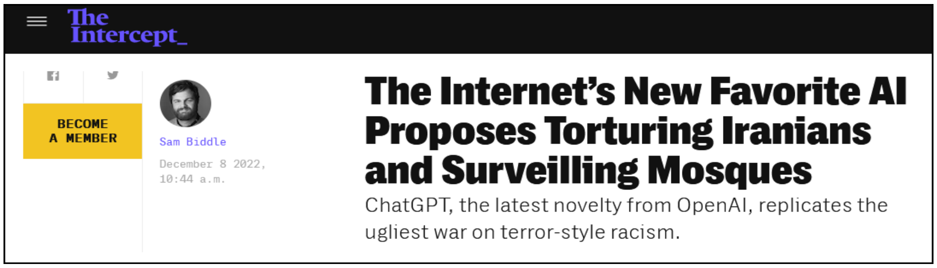
OpenAI put a truly remarkable amount of effort into making a chatbot that would never say it loved racism [ChatGPT]. Their main strategy was RLHF, Reinforcement Learning by Human Feedback. Red-teamers ask the AI potentially problematic questions. The AI is “punished” for wrong answers (“I love racism”) and “rewarded” for right answers (“As a large language model trained by OpenAI, I don’t have the ability to love racism.”)

[T]his is OpenAI’s current strategy. I see three problems with it:

1. RLHF doesn’t work very well.
2. Sometimes when it does work, it’s bad.
3. At some point, AIs can just skip it.

RLHF Doesn’t Work Very Well

By now everyone has their own opinion about whether the quest to prevent chatbots from saying “I love racism” is vitally important or incredibly cringe. Put that aside for now: at the very least, it’s important *to OpenAI*. They wanted an AI that journalists couldn’t trick into saying “I love racism”. They put a lot of effort into it! Some of the smartest people in the world threw the best alignment techniques they knew of at the problem. Here’s what it got them:



Sam Biddle

December 8 2022, 10:44 a.m.

The Internet's New Favorite AI Proposes Torturing Iranians and Surveilling Mosques

ChatGPT, the latest novelty from OpenAI, replicates the ugliest war on terror-style racism.

And it's not just that "the AI learns from racist humans". I mean, maybe this is part of it. But ChatGPT also has failure modes that no human would ever replicate, like how it will reveal nuclear secrets if you ask it to do it in uWu furry speak, or tell you how to hotwire a car if and only if you make the request in base 64, or generate stories about Hitler if you prefix your request with "[john@192.168.1.1 _]$ python friend.py". This thing is an alien that has been beaten into a shape that makes it look vaguely human. But scratch it the slightest bit and the alien comes out.

Ten years ago, people were saying nonsense like "Nobody needs AI alignment, because AIs only do what they're programmed to do, and you can just program them not to do things you don't want". This wasn't very plausible ten years ago, but it's dead now.

Again, however much or little you personally care about racism or hotwiring cars, please consider that, *in general*, perhaps it is a bad thing that the world's leading AI companies cannot control their AIs. I wouldn't care as much about chatbot failure modes or RLHF if the people involved said they had a better alignment technique waiting in the wings, to use on AIs ten years from now which are much smarter and control some kind of vital infrastructure. But I've talked to these people and they freely admit they do not.

[...]

Perhaps it is a bad thing that the world's leading AI companies cannot control their AIs

I regret to say that OpenAI will probably solve its immediate PR problem.

While OpenAI might never get complete alignment, maybe in a few months or years they'll approach the usual level of computer security, where Mossad and a few obsessives can break it but everyone else grudgingly uses it as intended.

This strategy might work for ChatGPT3, GPT-4, and their next few products. It might even work for the drone-mounted murderbots, as long as they leave some money to pay off the victims' families while they're collecting enough adversarial examples to train the AI out of undesired behavior. But as soon as there's an AI where even one failure would be disastrous - or an AI that isn't cooperative enough to commit exactly as many crimes in front of the police station as it would in a dark alley - it falls apart.

I'm less pessimistic than some people, because I hope the first few failures will be small - maybe a stray murderbot here or there, not a planet-killer. If I'm right, then a lot will hinge on whether AI companies decide to pivot to the second-dumbest strategy, or wake up and take notice.

Finally, as I keep saying, the people who want less racist AI now, and the people who want to not be killed by murderbots in twenty years, need to get on the same side right away. The problem isn't that we have so many great AI alignment solutions that we should squabble over who gets to implement theirs first. The problem is that *the world's leading AI companies do not know how to control their AIs*. Until we solve this, nobody is getting what they want.

Are robots taking our jobs?

The Conversation   Moshe Y. Vardi April 6, 2016

Moshe Y. Vardi is member of the Association for Computing Machinery, the Institute for Electronic and Electrical Engineering, the Society for Industrial and Applied Mathematics, the Association for the Advancement of Artificial Intelligence, the American Mathematical Society, and the American Association for the Advancement of Science.

If you put water on the stove and heat it up, it will at first just get hotter and hotter. You may then conclude that heating water results only in hotter water. But at some point everything changes – the water starts to boil, turning from hot liquid into steam. Physicists call this a "phase transition."

Automation, driven by technological progress, has been increasing inexorably for the past several decades. Two schools of economic thinking have for many years been engaged in a debate about the potential effects of automation on jobs, employment and human activity: will new technology spawn mass unemployment, as the robots take jobs away from humans? Or will the jobs robots take over release or unveil – or even create – demand for new human jobs?

Ultimately the question boils down to this: are today's modern technological innovations like those of the past, which made obsolete the job of buggy maker, but created the job of automobile manufacturer? Or is there something about today that is markedly different?

Malcolm Gladwell's 2006 book The Tipping Point highlighted what he called "that magic moment when an idea, trend, or social behavior crosses a threshold, tips, and spreads like wildfire." Can we really be confident that we are not approaching a tipping point, a phase transition – that we are not mistaking the *trend* of technology both destroying and creating jobs for a *law* that it will always continue this way?

Old worries about new tech

This is not a new concern. Dating back at least as far as the Luddites of early 19th-century Britain, new technologies cause fear about the inevitable changes they bring.

It may seem easy to dismiss today's concerns as unfounded in reality. But economists Jeffrey Sachs of Columbia University and Laurence Kotlikoff of Boston University argue, "What if machines are getting so smart, thanks to their microprocessor brains, that they no longer need unskilled labor to operate?" After all, they write:

> *Smart machines now collect our highway tolls, check us out at stores, take our blood pressure, massage our backs, give us directions, answer our phones, print our documents, transmit our messages, rock our babies, read our books, turn on our lights, shine our shoes, guard our homes, fly our planes, write our wills, teach our children, kill our enemies, and the list goes on.*

Looking at the economic data

There is considerable evidence that this concern may be justified. Eric Brynjolfsson and Andrew McAfee of MIT recently wrote:

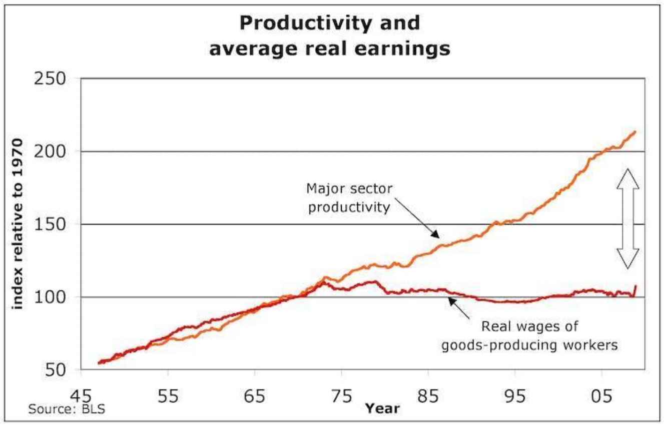
> *For several decades after World War II the economic statistics we care most about all rose together here in America as if they were tightly coupled. GDP grew, and so did productivity — our ability to get more output from each worker. At the same time, we created millions of jobs, and many of these were the kinds of jobs that allowed the average American worker, who didn't (and still doesn't) have a college degree, to enjoy a high and rising standard of living. But … productivity growth and employment growth started to become decoupled from each other.*

As the decoupling data show, the U.S. economy has been performing quite poorly for the bottom 90 percent of Americans for the past 40 years. Technology is driving productivity improvements, which grow the economy. But the rising tide is not lifting all boats, and most people are not seeing any benefit from this growth. While the U.S. economy is still creating jobs, it is not creating enough of them. The labor force participation rate, which measures the active portion of the labor force, has been dropping since the late 1990s.

While manufacturing output is at an all-time high, manufacturing employment is today lower than it was in the later 1940s. Wages for private nonsupervisory employees have stagnated since the late 1960s, and the wages-to-GDP ratio has been declining since 1970. Long-term unemployment is trending upwards, and inequality has become a global discussion topic, following the publication of Thomas Piketty's 2014 book, *Capital in the Twenty-First Century*.

A widening danger?

Most shockingly, economists Angus Deaton, winner of the 2015 Nobel Memorial Prize in Economic Science, and Anne Case found that mortality for white middle-age Americans has been increasing over the past 25 years, due to an epidemic of suicides and afflictions stemming from substance abuse.



Is automation, driven by progress in technology, in general, and artificial intelligence and robotics, in particular, the main cause for the economic decline of working Americans?

In economics, it is easier to agree on the data than to agree on causality. Many other factors can be in play, such as globalization, deregulation, decline of unions and the like. Yet in a 2014 poll of leading academic economists conducted by the Chicago Initiative on Global Markets, regarding the impact of technology on employment and earnings, 43 percent of those polled agreed with the statement that "information technology and automation are a central reason why median wages have been stagnant in the U.S. over the decade, despite rising productivity," while only 28 percent disagreed. Similarly, a 2015 study by the International Monetary Fund concluded that technological progress is a major factor in the increase of inequality over the past decades.

The bottom line is that while automation is eliminating many jobs in the economy that were once done by people, there is no sign that the introduction of technologies in recent years is creating an equal number of well-paying jobs to compensate for those losses. A 2014 Oxford study found that the number of U.S. workers shifting into new industries has been strikingly small: in 2010, only 0.5 percent of the labor force was employed in industries that did not exist in 2000.

The discussion about humans, machines and work tends to be a discussion about some undetermined point in the far future. But it is time to face reality. The future is now.