AI takeover

An **AI takeover** is a hypothetical scenario in which artificial intelligence (AI) becomes the dominant form of intelligence on Earth, with computers or robots effectively taking the control of the planet away from the human species. Possible scenarios include replacement of the entire human workforce, takeover by a superintelligent AI, and the popular notion of a robot uprising. Some public figures, such as Stephen Hawking and Elon Musk, have advocated research into precautionary measures to ensure future superintelligent machines remain under human control.[1]

Robots revolt in *R.U.R.*, a 1920 play

Contents

Types

Automation of the economy

The traditional consensus among economists has been that technological progress does not cause long-term unemployment. However, recent innovation in the fields of robotics and artificial intelligence has raised worries that human labor will become obsolete, leaving people in various sectors without jobs to earn a living,

leading to an economic crisis.[2][3][4][5] Many small and medium size businesses may also be driven out of business if they will not be able to afford or licence the latest robotic and AI technology, and may need to focus on areas or services that cannot easily be replaced for continued viability in the face of such technology.[6]

Technologies that may displace workers

Computer-integrated manufacturing

Computer-integrated manufacturing is the manufacturing approach of using computers to control the entire production process. This integration allows individual processes to exchange information with each other and initiate actions. Although manufacturing can be faster and less error-prone by the integration of computers, the main advantage is the ability to create automated manufacturing processes. Computer-integrated manufacturing is used in automotive, aviation, space, and ship building industries.

White-collar machines

The 21st century has seen a variety of skilled tasks partially taken over by machines, including translation, legal research and even low level journalism. Care work, entertainment, and other tasks requiring empathy, previously thought safe from automation, have also begun to be performed by robots.[7][8][9][10]

Autonomous cars

An autonomous car is a vehicle that is capable of sensing its environment and navigating without human input. Many such vehicles are being developed, but as of May 2017 automated cars permitted on public roads are not yet fully autonomous. They all require a human driver at the wheel who is ready at a moment's notice to take control of the vehicle. Among the main obstacles to widespread adoption of autonomous vehicles, are concerns about the resulting loss of driving-related jobs in the road transport industry. On March 18, 2018, the first human was killed by an autonomous vehicle in Tempe, Arizona by an Uber self-driving car.[11]

Eradication

Scientists such as Stephen Hawking are confident that superhuman artificial intelligence is physically possible, stating "there is no physical law precluding particles from being organised in ways that perform even more advanced computations than the arrangements of particles in human brains".[12][13] Scholars like Nick Bostrom debate how far off superhuman intelligence is, and whether it would actually pose a risk to mankind. A superintelligent machine would not necessarily be motivated by the same *emotional* desire to collect power that often drives human beings. However, a machine could be motivated to take over the world as a rational means toward attaining its ultimate goals; taking over the world would both increase its access to resources, and would help to prevent other agents from stopping the machine's plans. As an oversimplified example, a paperclip maximizer designed solely to create as many paperclips as possible would want to take over the world so that it can use all of the world's resources to create as many paperclips as possible, and, additionally, prevent humans from shutting it down or using those resources on things other than paperclips.[14]

In fiction

AI takeover is a common theme in science fiction. Fictional scenarios typically differ vastly from those hypothesized by researchers in that they involve an active conflict between humans and an AI or robots with anthropomorphic motives who see them as a threat or otherwise have active desire to fight humans, as opposed to the researchers' concern of an AI that rapidly exterminates humans as a byproduct of pursuing arbitrary goals.[15] This theme is at least as old as Karel Čapek's *R. U. R.*, which introduced the word *robot* to the global lexicon in 1921,[16] and can even be glimpsed in Mary Shelley's *Frankenstein* (published in 1818), as Victor ponders whether, if he grants his monster's request and makes him a wife, they would reproduce and their kind would destroy humanity.[17]

The word "robot" from *R.U.R.* comes from the Czech word, robota, meaning laborer or serf. The 1920 play was a protest against the rapid growth of technology, featuring manufactured "robots" with increasing capabilities who eventually revolt.[18] HAL 9000 (1968) and the original Terminator (1984) are two iconic examples of hostile AI in pop culture.[19]

Contributing factors

Advantages of superhuman intelligence over humans

Nick Bostrom and others have expressed concern that an AI with the abilities of a competent artificial intelligence researcher would be able to modify its own source code and increase its own intelligence. If its self-reprogramming leads to its getting even better at being able to reprogram itself, the result could be a recursive intelligence explosion where it would rapidly leave human intelligence far behind. Bostrom defines a superintelligence as "any intellect that greatly exceeds the cognitive performance of humans in virtually all domains of interest", and enumerates some advantages a superintelligence would have if it chose to compete against humans:[15][20]

- Technology research: A machine with superhuman scientific research abilities would be able to beat the human research community to milestones such as nanotechnology or advanced biotechnology. If the advantage becomes sufficiently large (for example, due to a sudden intelligence explosion), an AI takeover becomes trivial. For example, a superintelligent AI might design self-replicating bots that initially escape detection by diffusing throughout the world at a low concentration. Then, at a prearranged time, the bots multiply into nanofactories that cover every square foot of the Earth, producing nerve gas or deadly target-seeking mini-drones.
- Strategizing: A superintelligence might be able to simply outwit human opposition.
- Social manipulation: A superintelligence might be able to recruit human support,[15] or covertly incite a war between humans.[21]
- Economic productivity: As long as a copy of the AI could produce more economic wealth than the cost of its hardware, individual humans would have an incentive to voluntarily allow the Artificial General Intelligence (AGI) to run a copy of itself on their systems.
- Hacking: A superintelligence could find new exploits in computers connected to the Internet, and spread copies of itself onto those systems, or might steal money to finance its plans.

Sources of AI advantage

According to Bostrom, a computer program that faithfully emulates a human brain, or that otherwise runs algorithms that are equally powerful as the human brain's algorithms, could still become a "speed superintelligence" if it can think many orders of magnitude faster than a human, due to being made of silicon rather than flesh, or due to optimization focusing on increasing the speed of the AGI. Biological neurons

operate at about 200 Hz, whereas a modern microprocessor operates at a speed of about 2,000,000,000 Hz. Human axons carry action potentials at around 120 m/s, whereas computer signals travel near the speed of light.[15]

A network of human-level intelligences designed to network together and share complex thoughts and memories seamlessly, able to collectively work as a giant unified team without friction, or consisting of trillions of human-level intelligences, would become a "collective superintelligence".[15]

More broadly, any number of qualitative improvements to a human-level AGI could result in a "quality superintelligence", perhaps resulting in an AGI as far above us in intelligence as humans are above non-human apes. The number of neurons in a human brain is limited by cranial volume and metabolic constraints, while the number of processors in a supercomputer can be indefinitely expanded. An AGI need not be limited by human constraints on working memory, and might therefore be able to intuitively grasp more complex relationships than humans can. An AGI with specialized cognitive support for engineering or computer programming would have an advantage in these fields, compared with humans who evolved no specialized mental modules to specifically deal with those domains. Unlike humans, an AGI can spawn copies of itself and tinker with its copies' source code to attempt to further improve its algorithms.[15]

Possibility of unfriendly AI preceding friendly AI

Is strong AI inherently dangerous?

A significant problem is that unfriendly artificial intelligence is likely to be much easier to create than friendly AI. While both require large advances in recursive optimisation process design, friendly AI also requires the ability to make goal structures invariant under self-improvement (or the AI could transform itself into something unfriendly) and a goal structure that aligns with human values and does not automatically destroy the entire human race. An unfriendly AI, on the other hand, can optimize for an arbitrary goal structure, which does not need to be invariant under self-modification.[22]

The sheer complexity of human value systems makes it very difficult to make AI's motivations human-friendly.[15][23] Unless moral philosophy provides us with a flawless ethical theory, an AI's utility function could allow for many potentially harmful scenarios that conform with a given ethical framework but not "common sense". According to Eliezer Yudkowsky, there is little reason to suppose that an artificially designed mind would have such an adaptation.[24]

Odds of conflict

Many scholars, including as evolutionary psychologist Steven Pinker, argue that a superintelligent machine is likely to coexist peacefully with humans.[25]

The fear of cybernetic revolt is often based on interpretations of humanity's history, which is rife with incidents of enslavement and genocide. Such fears stem from a belief that competitiveness and aggression are necessary in any intelligent being's goal system. However, such human competitiveness stems from the evolutionary background to our intelligence, where the survival and reproduction of genes in the face of human and non-human competitors was the central goal.[26] According to AI researcher Steve Omohundro, an arbitrary intelligence could have arbitrary goals: there is no particular reason that an artificially intelligent machine (not sharing humanity's evolutionary context) would be hostile—or friendly—unless its creator programs it to be such and it is not inclined or capable of modifying its programming. But the question remains: what would

happen if AI systems could interact and evolve (evolution in this context means self-modification or selection and reproduction) and need to compete over resources, would that create goals of self-preservation? AI's goal of self-preservation could be in conflict with some goals of humans.[27]

Many scholars dispute the likelihood of unanticipated cybernetic revolt as depicted in science fiction such as *The Matrix*, arguing that it is more likely that any artificial intelligence powerful enough to threaten humanity would probably be programmed not to attack it. Pinker acknowledges the possibility of deliberate "bad actors", but states that in the absence of bad actors, unanticipated accidents are not a significant threat; Pinker argues that a culture of engineering safety will prevent AI researchers from unleashing malign superintelligence on accident.[25] In contrast, Yudkowsky argues that humanity is less likely to be threatened by deliberately aggressive AIs than by AIs which were programmed such that their goals are unintentionally incompatible with human survival or well-being (as in the film *I, Robot* and in the short story "The Evitable Conflict"). Omohundro suggests that present-day automation systems are not designed for safety and that AIs may blindly optimize narrow utility functions (say, playing chess at all costs), leading them to seek self-preservation and elimination of obstacles, including humans who might turn them off.[28]

Precautions

The **AI control problem** is the issue of how to build a superintelligent agent that will aid its creators, and avoid inadvertently building a superintelligence that will harm its creators. Some scholars argue that solutions to the control problem might also find applications in existing non-superintelligent AI.[29]

Major approaches to the control problem include *alignment*, which aims to align AI goal systems with human values, and *capability control*, which aims to reduce an AI system's capacity to harm humans or gain control. An example of "capability control" is to research whether a superintelligence AI could be successfully confined in an "AI box". According to Bostrom, such capability control proposals are not reliable or sufficient to solve the control problem in the long term, but may potentially act as valuable supplements to alignment efforts.[15]

Warnings

Physicist Stephen Hawking, Microsoft founder Bill Gates and SpaceX founder Elon Musk have expressed concerns about the possibility that AI could develop to the point that humans could not control it, with Hawking theorizing that this could "spell the end of the human race".[30] Stephen Hawking said in 2014 that "Success in creating AI would be the biggest event in human history. Unfortunately, it might also be the last, unless we learn how to avoid the risks." Hawking believed that in the coming decades, AI could offer "incalculable benefits and risks" such as "technology outsmarting financial markets, out-inventing human researchers, out-manipulating human leaders, and developing weapons we cannot even understand." In January 2015, Nick Bostrom joined Stephen Hawking, Max Tegmark, Elon Musk, Lord Martin Rees, Jaan Tallinn, and numerous AI researchers, in signing the Future of Life Institute's open letter speaking to the potential risks and benefits associated with artificial intelligence. The signatories

> …believe that research on how to make AI systems robust and beneficial is both important and timely, and that there are concrete research directions that can be pursued today.[31][32]

See also

- Artificial intelligence arms race

- Autonomous robot
  - Industrial robot
  - Mobile robot
  - Self-replicating machine
- Effective altruism
- Existential risk from artificial general intelligence
- Future of Humanity Institute
- Global catastrophic risk (existential risk)
- Government by algorithm
- Machine ethics
- Machine learning/Deep learning
- Nick Bostrom
- Outline of transhumanism
- Self-replication
- Technological singularity
  - Intelligence explosion
  - Superintelligence
    - *Superintelligence: Paths, Dangers, Strategies*

References

1. Lewis, Tanya (2015-01-12). "*Don't Let Artificial Intelligence Take Over, Top Scientists Warn*" (http://www.livescience.com/49419-artificial-intelligence-dangers-letter.html). *LiveScience*. Purch. Retrieved October 20, 2015. "Stephen Hawking, Elon Musk and dozens of other top scientists and technology leaders have signed a letter warning of the potential dangers of developing artificial intelligence (AI)."
2. Lee, Kai-Fu (2017-06-24). "The Real Threat of Artificial Intelligence" (https://www.nytimes.com/2017/06/24/opinion/sunday/artificial-intelligence-economic-inequality.html). *The New York Times*. Retrieved 2017-08-15. "These tools can outperform human beings at a given task. This kind of A.I. is spreading to thousands of domains, and as it does, it will eliminate many jobs."
3. Larson, Nina (2017-06-08). "AI 'good for the world'... says ultra-lifelike robot" (https://phys.org/news/2017-06-ai-good-world-ultra-lifelike-robot.html). *Phys.org*. Phys.org. Retrieved 2017-08-15. "Among the feared consequences of the rise of the robots is the growing impact they will have on human jobs and economies."
4. Santini, Jean-Louis (2016-02-14). "Intelligent robots threaten millions of jobs" (https://phys.org/news/2016-02-intelligent-robots-threaten-millions-jobs.html#nRlv). *Phys.org*. Phys.org. Retrieved 2017-08-15. ""We are approaching a time when machines will be able to outperform humans at almost any task," said Moshe Vardi, director of the Institute for Information Technology at Rice University in Texas."
5. Williams-Grut, Oscar (2016-02-15). "Robots will steal your job: How AI could increase unemployment and inequality" (http://www.businessinsider.com/robots-will-steal-your-job-citi-ai-increase-unemployment-inequality-2016-2?r=UK&IR=T). *Businessinsider.com*. Business Insider. Retrieved 2017-08-15. "Top computer scientists in the US warned that the rise of artificial intelligence (AI) and robots in the workplace could cause mass unemployment and dislocated economies, rather than simply unlocking productivity gains and freeing us all up to watch TV and play sports."
6. "How can SMEs prepare for the rise of the robots?" (https://web.archive.org/web/20171018073852/http://www.leanstaff.co.uk/robot-apocalypse/). *LeanStaff*. 2017-10-17. Archived from the original (http://www.leanstaff.co.uk/robot-apocalypse/) on 2017-10-18. Retrieved 2017-10-17.

7. Skidelsky, Robert (2013-02-19). "Rise of the robots: what will the future of work look like?" (https://www.theguardian.com/business/2013/feb/19/rise-of-robots-future-of-work). London: The Guardian. Retrieved 14 July 2015.
8. Bria, Francesca (February 2016). "The robot economy may already have arrived" (https://www.opendemocracy.net/can-europe-make-it/francesca-bria/robot-economy-full-automation-work-future). openDemocracy. Retrieved 20 May 2016.
9. Srnicek, Nick (March 2016). "4 Reasons Why Technological Unemployment Might Really Be Different This Time" (https://web.archive.org/web/20160625161447/http://wire.novaramedia.com/2015/03/4-reasons-why-technological-unemployment-might-really-be-different-this-time/). novara wire. Archived from the original (http://wire.novaramedia.com/2015/03/4-reasons-why-technological-unemployment-might-really-be-different-this-time/) on 25 June 2016. Retrieved 20 May 2016.
10. Erik Brynjolfsson and Andrew McAfee (2014). "*passim*, see esp Chpt. 9". *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*. W. W. Norton & Company. ISBN 978-0393239355.
11. Wakabayashi, Daisuke (March 19, 2018). "Self-Driving Uber Car Kills Pedestrian in Arizona, Where Robots Roam" (https://www.nytimes.com/2018/03/19/technology/uber-driverless-fatality.html). *New York Times*.
12. Stephen Hawking; Stuart Russell; Max Tegmark; Frank Wilczek (1 May 2014). "Stephen Hawking: 'Transcendence looks at the implications of artificial intelligence - but are we taking AI seriously enough?' " (https://www.independent.co.uk/news/science/stephen-hawking-transcendence-looks-at-the-implications-of-artificial-intelligence-but-are-we-taking-9313474.html). *The Independent*. Retrieved 1 April 2016.
13. Vincent C. Müller and Nick Bostrom. "Future progress in artificial intelligence: A survey of expert opinion." In Fundamental issues of artificial intelligence, pp. 555-572. Springer, Cham, 2016. "AI systems will... reach overall human ability... very likely (with 90% probability) by 2075. From reaching human ability, it will move on to superintelligence within 30 years (75%)... So, (most of the AI experts responding to the surveys) think that superintelligence is likely to come in a few decades..."
14. Bostrom, Nick. "The superintelligent will: Motivation and instrumental rationality in advanced artificial agents." Minds and Machines 22.2 (2012): 71-85.
15. Bostrom, Nick. *Superintelligence: Paths, Dangers, Strategies*.
16. "The Origin Of The Word 'Robot' " (https://www.sciencefriday.com/segments/the-origin-of-the-word-robot/). *Science Friday (public radio)*. 22 April 2011. Retrieved 30 April 2020.
17. Botkin-Kowacki, Eva (28 October 2016). "A female Frankenstein would lead to humanity's extinction, say scientists" (https://www.csmonitor.com/Science/2016/1028/A-female-Frankenstein-would-lead-to-humanity-s-extinction-say-scientists). *Christian Science Monitor*. Retrieved 30 April 2020.
18. Hockstein, N. G.; Gourin, C. G.; Faust, R. A.; Terris, D. J. (17 March 2007). "A history of robots: from science fiction to surgical robotics" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4247417). *Journal of Robotic Surgery*. **1** (2): 113–118. doi:10.1007/s11701-007-0021-2 (https://doi.org/10.1007%2Fs11701-007-0021-2). PMC 4247417 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4247417). PMID 25484946 (https://pubmed.ncbi.nlm.nih.gov/25484946).
19. Hellmann, Melissa (21 September 2019). "AI 101: What is artificial intelligence and where is it going?" (https://www.seattletimes.com/business/technology/ai-101-what-is-artificial-intelligence-and-where-is-it-going/). *The Seattle Times*. Retrieved 30 April 2020.
20. Babcock, James; Krámar, János; Yampolskiy, Roman V. (2019). "Guidelines for Artificial Intelligence Containment": 90–112. doi:10.1017/9781108616188.008 (https://doi.org/10.1017%2F9781108616188.008).
21. Baraniuk, Chris (23 May 2016). "Checklist of worst-case scenarios could help prepare for evil AI" (https://www.newscientist.com/article/2089606-checklist-of-worst-case-scenarios-could-help-prepare-for-evil-ai/). *New Scientist*. Retrieved 21 September 2016.

22. Yudkowsky, Eliezer S. (May 2004). "Coherent Extrapolated Volition" (https://web.archive.org/web/20120615203944/http://singinst.org/upload/CEV.html). Singularity Institute for Artificial Intelligence. Archived from the original (http://singinst.org/upload/CEV.html) on 2012-06-15.
23. Muehlhauser, Luke; Helm, Louie (2012). "Intelligence Explosion and Machine Ethics" (https://intelligence.org/files/IE-ME.pdf) (PDF). *Singularity Hypotheses: A Scientific and Philosophical Assessment*. Springer.
24. Yudkowsky, Eliezer (2011). "Complex Value Systems in Friendly AI". **6830**: 388–393. doi:10.1007/978-3-642-22887-2_48 (https://doi.org/10.1007%2F978-3-642-22887-2_48). ISSN 0302-9743 (https://www.worldcat.org/issn/0302-9743).
25. Pinker, Steven (13 February 2018). "We're told to fear robots. But why do we think they'll turn on us?" (https://www.popsci.com/robot-uprising-enlightenment-now/). *Popular Science*. Retrieved 8 June 2020.
26. *Creating a New Intelligent Species: Choices and Responsibilities for Artificial Intelligence Designers (http://www.singinst.org/ourresearch/presentations/) Archived (https://web.archive.org/web/20070206060938/http://www.singinst.org/ourresearch/presentations/) February 6, 2007, at the Wayback Machine* - Singularity Institute for Artificial Intelligence, 2005
27. Omohundro, Stephen M. (June 2008). *The basic AI drives* (https://selfawaresystems.files.wordpress.com/2008/01/ai_drives_final.pdf) (PDF). Artificial General Intelligence 2008. pp. 483–492.
28. Tucker, Patrick (17 Apr 2014). "Why There Will Be A Robot Uprising" (http://www.defenseone.com/technology/2014/04/why-there-will-be-robot-uprising/82783/). Defense One. Retrieved 15 July 2014.
29. "Google developing kill switch for AI" (https://www.bbc.com/news/technology-36472140). *BBC News*. 8 June 2016. Retrieved 7 June 2020.
30. Rawlinson, Kevin (29 January 2015). "Microsoft's Bill Gates insists AI is a threat" (https://www.bbc.co.uk/news/31047780). BBC News. Retrieved 30 January 2015.
31. "The Future of Life Institute Open Letter" (http://futureoflife.org/ai-open-letter). The Future of Life Institute. Retrieved 29 March 2019.
32. Bradshaw, Tim (11 January 2015). "Scientists and investors warn on AI" (https://www.ft.com/cms/s/0/3d2c2f12-99e9-11e4-93c1-00144feabdc0.html#axzz3TNL9lxJV). The Financial Times. Retrieved 4 March 2015.

External links

- Automation, not domination: How robots will take over our world (http://robohub.org/automation-not-domination-how-robots-will-take-over-our-world/) (a positive outlook of robot and AI integration into society)
- Machine Intelligence Research Institute (http://www.intelligence.org/): official MIRI (formerly Singularity Institute for Artificial Intelligence) website
- Lifeboat Foundation AIShield (http://lifeboat.com/ex/ai.shield/) (To protect against unfriendly AI)
- Ted talk: Can we build AI without losing control over it? (https://www.youtube.com/watch?v=R_sSpPyruj0)