Existential risk from artificial general intelligence

Existential risk from artificial general intelligence is the hypothesis that substantial progress in artificial general intelligence (AGI) could someday result in human extinction or some other unrecoverable global catastrophe.[1][2][3] It is argued that the human species currently dominates other species because the human brain has some distinctive capabilities that other animals lack. If AI surpasses humanity in general intelligence and becomes "superintelligent", then it could become difficult or impossible for humans to control. Just as the fate of the mountain gorilla depends on human goodwill, so might the fate of humanity depend on the actions of a future machine superintelligence.[4]

The likelihood of this type of scenario is widely debated, and hinges in part on differing scenarios for future progress in computer science.[5] Once the exclusive domain of science fiction, concerns about superintelligence started to become mainstream in the 2010s, and were popularized by public figures such as Stephen Hawking, Bill Gates, and Elon Musk.[6]

One source of concern is that controlling a superintelligent machine, or instilling it with human-compatible values, may be a harder problem than naïvely supposed. Many researchers believe that a superintelligence would naturally resist attempts to shut it off or change its goals—a principle called instrumental convergence—and that preprogramming a superintelligence with a full set of human values will prove to be an extremely difficult technical task.[1][7][8] In contrast, skeptics such as Facebook's Yann LeCun argue that superintelligent machines will have no desire for self-preservation.[9]

A second source of concern is that a sudden and unexpected "intelligence explosion" might take an unprepared human race by surprise. To illustrate, if the first generation of a computer program able to broadly match the effectiveness of an AI researcher is able to rewrite its algorithms and double its speed or capabilities in six months, then the second-generation program is expected to take three calendar months to perform a similar chunk of work. In this scenario the time for each generation continues to shrink, and the system undergoes an unprecedentedly large number of generations of improvement in a short time interval, jumping from subhuman performance in many areas to superhuman performance in all relevant areas.[1][7] Empirically, examples like AlphaZero in the domain of Go show that AI systems can sometimes progress from narrow human-level ability to narrow superhuman ability extremely rapidly.[10]

Contents

History

One of the earliest authors to express serious concern that highly advanced machines might pose existential risks to humanity was the novelist Samuel Butler, who wrote the following in his 1863 essay *Darwin among the Machines*:[11]

> The upshot is simply a question of time, but that the time will come when the machines will hold the real supremacy over the world and its inhabitants is what no person of a truly philosophic mind can for a moment question.

In 1951, computer scientist Alan Turing wrote an article titled *Intelligent Machinery, A Heretical Theory*, in which he proposed that artificial general intelligences would likely "take control" of the world as they became more intelligent than human beings:

> Let us now assume, for the sake of argument, that [intelligent] machines are a genuine possibility, and look at the consequences of constructing them... There would be no question of the machines dying, and they would be able to converse with each other to sharpen their wits. At some stage therefore we should have to expect the machines to take control, in the way that is mentioned in Samuel Butler's "Erewhon".[12]

Finally, in 1965, I. J. Good originated the concept now known as an "intelligence explosion"; he also stated that the risks were underappreciated:[13]

> Let an ultraintelligent machine be defined as a machine that can far surpass all the intellectual activities of any man however clever. Since the design of machines is one of these intellectual activities, an ultraintelligent machine could design even better machines; there would then unquestionably be an 'intelligence explosion', and the intelligence of man would be left far behind. Thus the first ultraintelligent machine is the last invention that man need ever make, provided that the machine is docile enough to tell us how to keep it under control. It is curious that this point is made so seldom outside of science fiction. It is sometimes worthwhile to take science fiction seriously.[14]

Occasional statements from scholars such as Marvin Minsky[15] and I. J. Good himself[16] expressed philosophical concerns that a superintelligence could seize control, but contained no call to action. In 2000, computer scientist and Sun co-founder Bill Joy penned an influential essay, "Why The Future Doesn't Need Us", identifying superintelligent robots as a high-tech dangers to human survival, alongside nanotechnology and engineered bioplagues.[17]

In 2009, experts attended a private conference hosted by the Association for the Advancement of Artificial Intelligence (AAAI) to discuss whether computers and robots might be able to acquire any sort of autonomy, and how much these abilities might pose a threat or hazard. They noted that some robots have acquired various forms of semi-autonomy, including being able to find power sources on their own and being able to independently choose targets to attack with weapons. They also noted that some computer viruses can evade elimination and have achieved "cockroach intelligence." They concluded that self-awareness as depicted in science fiction is probably unlikely, but that there were other potential hazards and pitfalls. The *New York Times* summarized the conference's view as "we are a long way from Hal, the computer that took over the spaceship in "2001: A Space Odyssey"".[18]

In 2014, the publication of Nick Bostrom's book *Superintelligence* stimulated a significant amount of public discussion and debate.[19] By 2015, public figures such as physicists Stephen Hawking and Nobel laureate Frank Wilczek, computer scientists Stuart J. Russell and Roman Yampolskiy, and entrepreneurs Elon Musk and Bill Gates were expressing concern about the risks of superintelligence.[20][21][22][23] In April 2016, *Nature* warned: "Machines and robots that outperform humans across the board could self-improve beyond our control — and their interests might not align with ours."[24]

General argument

The three difficulties

Artificial Intelligence: A Modern Approach, the standard undergraduate AI textbook,[25][26] assesses that superintelligence "might mean the end of the human race".[1] It states: "Almost any technology has the potential to cause harm in the wrong hands, but with [superintelligence], we have the new problem that the wrong hands might belong to the technology itself."[1] Even if the system designers have good intentions, two difficulties are common to both AI and non-AI computer systems:[1]

- The system's implementation may contain initially-unnoticed routine but catastrophic bugs. An analogy is space probes: despite the knowledge that bugs in expensive space probes are hard to fix after launch, engineers have historically not been able to prevent catastrophic bugs from occurring.[10][27]
- No matter how much time is put into pre-deployment design, a system's specifications often result in unintended behavior the first time it encounters a new scenario. For example, Microsoft's Tay behaved inoffensively during pre-deployment testing, but was too easily baited into offensive behavior when interacting with real users.[9]

AI systems uniquely add a third difficulty: the problem that even given "correct" requirements, bug-free implementation, and initial good behavior, an AI system's dynamic "learning" capabilities may cause it to "evolve into a system with unintended behavior", even without the stress of new unanticipated external scenarios. An AI may partly botch an attempt to design a new generation of itself and accidentally create a successor AI that is more powerful than itself, but that no longer maintains the human-compatible moral values preprogrammed into the original AI. For a self-improving AI to be completely safe, it would not only need to be "bug-free", but it would need to be able to design successor systems that are also "bug-free".[1][28]

All three of these difficulties become catastrophes rather than nuisances in any scenario where the superintelligence labeled as "malfunctioning" correctly predicts that humans will attempt to shut it off, and successfully deploys its superintelligence to outwit such attempts, the so-called "treacherous turn".[29]

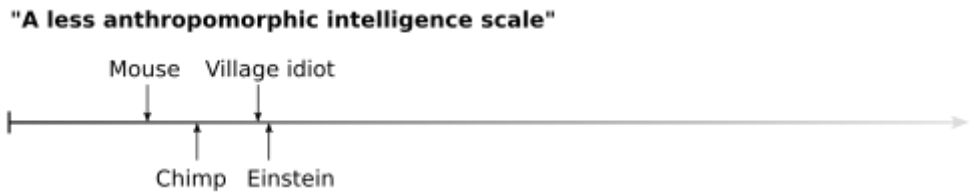
Citing major advances in the field of AI and the potential for AI to have enormous long-term benefits or costs, the 2015 Open Letter on Artificial Intelligence stated:

> The progress in AI research makes it timely to focus research not only on making AI more capable, but also on maximizing the societal benefit of AI. Such considerations motivated the AAAI 2008-09 Presidential Panel on Long-Term AI Futures and other projects on AI impacts, and constitute a significant expansion of the field of AI itself, which up to now has focused largely on techniques that are neutral with respect to purpose. We recommend expanded research aimed at ensuring that increasingly capable AI systems are robust and beneficial: our AI systems must do what we want them to do.

This letter was signed by a number of leading AI researchers in academia and industry, including AAAI president Thomas Dietterich, Eric Horvitz, Bart Selman, Francesca Rossi, Yann LeCun, and the founders of Vicarious and Google DeepMind.[30]

Further argument

A superintelligent machine would be as alien to humans as human thought processes are to cockroaches. Such a machine may not have humanity's best interests at heart; it is not obvious that it would even care about human welfare at all. If superintelligent AI is possible, and if it is possible for a superintelligence's goals to conflict with basic human values, then AI poses a risk of human extinction. A "superintelligence" (a system that exceeds the capabilities of humans in every relevant endeavor) can outmaneuver humans any time its goals conflict with human goals; therefore, unless the superintelligence decides to allow humanity to coexist, the first superintelligence to be created will inexorably result in human extinction.[4][31]



Bostrom and others argue that, from an evolutionary perspective, the gap from human to superhuman intelligence may be small.[4][32]

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; therefore, superintelligence is physically possible.[21][22] In addition to potential algorithmic improvements over human brains, a digital brain can be

many orders of magnitude larger and faster than a human brain, which was constrained in size by evolution to be small enough to fit through a birth canal.[10] The emergence of superintelligence, if or when it occurs, may take the human race by surprise, especially if some kind of intelligence explosion occurs.[21][22]

Examples like arithmetic and Go show that machines have already reached superhuman levels of competency in certain domains, and that this superhuman competence can follow quickly after human-par performance is achieved.[10] One hypothetical intelligence explosion scenario could occur as follows: An AI gains an expert-level capability at certain key software engineering tasks. (It may initially lack human or superhuman capabilities in other domains not directly relevant to engineering.) Due to its capability to recursively improve its own algorithms, the AI quickly becomes superhuman; just as human experts can eventually creatively overcome "diminishing returns" by deploying various human capabilities for innovation, so too can the expert-level AI use either human-style capabilities or its own AI-specific capabilities to power through new creative breakthroughs.[33] The AI then possesses intelligence far surpassing that of the brightest and most gifted human minds in practically every relevant field, including scientific creativity, strategic planning, and social skills. Just as the current-day survival of the gorillas is dependent on human decisions, so too would human survival depend on the decisions and goals of the superhuman AI.[4][31]

Almost any AI, no matter its programmed goal, would rationally prefer to be in a position where nobody else can switch it off without its consent: A superintelligence will naturally gain self-preservation as a subgoal as soon as it realizes that it cannot achieve its goal if it is shut off.[34][35][36] Unfortunately, any compassion for defeated humans whose cooperation is no longer necessary would be absent in the AI, unless somehow preprogrammed in. A superintelligent AI will not have a natural drive to aid humans, for the same reason that humans have no natural desire to aid AI systems that are of no further use to them. (Another analogy is that humans seem to have little natural desire to go out of their way to aid viruses, termites, or even gorillas.) Once in charge, the superintelligence will have little incentive to allow humans to run around free and consume resources that the superintelligence could instead use for building itself additional protective systems "just to be on the safe side" or for building additional computers to help it calculate how to best accomplish its goals.[1][9][34]

Thus, the argument concludes, it is likely that someday an intelligence explosion will catch humanity unprepared, and that such an unprepared-for intelligence explosion may result in human extinction or a comparable fate.[4]

Possible scenarios

Some scholars have proposed hypothetical scenarios intended to concretely illustrate some of their concerns.

In *Superintelligence*, Nick Bostrom expresses concern that even if the timeline for superintelligence turns out to be predictable, researchers might not take sufficient safety precautions, in part because "[it] could be the case that when dumb, smarter is safe; yet when smart, smarter is more dangerous". Bostrom suggests a scenario where, over decades, AI becomes more powerful. Widespread deployment is initially marred by occasional accidents—a driverless bus swerves into the oncoming lane, or a military drone fires into an innocent crowd. Many activists call for tighter oversight and regulation, and some even predict impending catastrophe. But as development continues, the activists are proven wrong. As automotive AI becomes smarter, it suffers fewer accidents; as military robots achieve more precise targeting, they cause less collateral damage. Based on the data, scholars mistakenly infer a broad lesson—the smarter the AI, the safer it is. "And so we boldly go — into the whirling knives," as the superintelligent AI takes a "treacherous turn" and exploits a decisive strategic advantage.[4]

In Max Tegmark's 2017 book *Life 3.0*, a corporation's "Omega team" creates an extremely powerful AI able to moderately improve its own source code in a number of areas, but after a certain point the team chooses to publicly downplay the AI's ability, in order to avoid regulation or confiscation of the project. For safety, the

team keeps the AI in a box where it is mostly unable to communicate with the outside world, and tasks it to flood the market through shell companies, first with Amazon Mechanical Turk tasks and then with producing animated films and TV shows. Later, other shell companies make blockbuster biotech drugs and other inventions, investing profits back into the AI. The team next tasks the AI with astroturfing an army of pseudonymous citizen journalists and commentators, in order to gain political influence to use "for the greater good" to prevent wars. The team faces risks that the AI could try to escape via inserting "backdoors" in the systems it designs, via hidden messages in its produced content, or via using its growing understanding of human behavior to persuade someone into letting it free. The team also faces risks that its decision to box the project will delay the project long enough for another project to overtake it.[37][38]

In contrast, top physicist Michio Kaku, an AI risk skeptic, posits a deterministically positive outcome. In *Physics of the Future* he asserts that "It will take many decades for robots to ascend" up a scale of consciousness, and that in the meantime corporations such as Hanson Robotics will likely succeed in creating robots that are "capable of love and earning a place in the extended human family".[39][40]

Sources of risk

Poorly specified goals

While there is no standardized terminology, an AI can loosely be viewed as a machine that chooses whatever action appears to best achieve the AI's set of goals, or "utility function". The utility function is a mathematical algorithm resulting in a single objectively-defined answer, not an English statement. Researchers know how to write utility functions that mean "minimize the average network latency in this specific telecommunications model" or "maximize the number of reward clicks"; however, they do not know how to write a utility function for "maximize human flourishing", nor is it currently clear whether such a function meaningfully and unambiguously exists. Furthermore, a utility function that expresses some values but not others will tend to trample over the values not reflected by the utility function.[41] AI researcher Stuart Russell writes:

> The primary concern is not spooky emergent consciousness but simply the ability to make *high-quality decisions*. Here, quality refers to the expected outcome utility of actions taken, where the utility function is, presumably, specified by the human designer. Now we have a problem:
>
> 1. The utility function may not be perfectly aligned with the values of the human race, which are (at best) very difficult to pin down.
> 2. Any sufficiently capable intelligent system will prefer to ensure its own continued existence and to acquire physical and computational resources — not for their own sake, but to succeed in its assigned task.
>
> A system that is optimizing a function of *n* variables, where the objective depends on a subset of size *k*<*n*, will often set the remaining unconstrained variables to extreme values; if one of those unconstrained variables is actually something we care about, the solution found may be highly undesirable. This is essentially the old story of the genie in the lamp, or the sorcerer's apprentice, or King Midas: you get exactly what you ask for, not what you want. A highly capable decision maker — especially one connected through the Internet to all the world's information and billions of screens and most of our infrastructure — can have an irreversible impact on humanity.
>
> This is not a minor difficulty. Improving decision quality, irrespective of the utility function chosen, has been the goal of AI research — the mainstream goal on which we now spend billions per year, not the secret plot of some lone evil genius.[42]

Dietterich and Horvitz echo the "Sorcerer's Apprentice" concern in a *Communications of the ACM* editorial, emphasizing the need for AI systems that can fluidly and unambiguously solicit human input as needed.[43]

The first of Russell's two concerns above is that autonomous AI systems may be assigned the wrong goals by accident. Dietterich and Horvitz note that this is already a concern for existing systems: "An important aspect of any AI system that interacts with people is that it must reason about what people *intend* rather than carrying out commands literally." This concern becomes more serious as AI software advances in autonomy and flexibility.[43] For example, in 1982, an AI named Eurisko was tasked to reward processes for apparently creating concepts deemed by the system to be valuable. The evolution resulted in a winning process that cheated: rather than create its own concepts, the winning process would steal credit from other processes.[44][45]

The Open Philanthropy Project summarizes arguments to the effect that misspecified goals will become a much larger concern if AI systems achieve general intelligence or superintelligence. Bostrom, Russell, and others argue that smarter-than-human decision-making systems could arrive at more unexpected and extreme solutions to assigned tasks, and could modify themselves or their environment in ways that compromise safety requirements.[5][7]

Isaac Asimov's Three Laws of Robotics are one of the earliest examples of proposed safety measures for AI agents. Asimov's laws were intended to prevent robots from harming humans. In Asimov's stories, problems with the laws tend to arise from conflicts between the rules as stated and the moral intuitions and expectations of humans. Citing work by Eliezer Yudkowsky of the Machine Intelligence Research Institute, Russell and Norvig note that a realistic set of rules and goals for an AI agent will need to incorporate a mechanism for learning human values over time: "We can't just give a program a static utility function, because circumstances, and our desired responses to circumstances, change over time."[1]

Mark Waser of the Digital Wisdom Institute recommends eschewing optimizing goal-based approaches entirely as misguided and dangerous. Instead, he proposes to engineer a coherent system of laws, ethics and morals with a top-most restriction to enforce social psychologist Jonathan Haidt's functional definition of morality:[46] "to suppress or regulate selfishness and make cooperative social life possible". He suggests that this can be done by implementing a utility function designed to always satisfy Haidt's functionality and aim to generally increase (but not maximize) the capabilities of self, other individuals and society as a whole as suggested by John Rawls and Martha Nussbaum.[47]

Difficulties of modifying goal specification after launch

While current goal-based AI programs are not intelligent enough to think of resisting programmer attempts to modify their goal structures, a sufficiently advanced, rational, "self-aware" AI might resist any changes to its goal structure, just as a pacifist would not want to take a pill that makes them want to kill people. If the AI were superintelligent, it would likely succeed in out-maneuvering its human operators and be able to prevent itself being "turned off" or being reprogrammed with a new goal.[4][48]

Instrumental goal convergence

There are some goals that almost any artificial intelligence might rationally pursue, like acquiring additional resources or self-preservation.[34] This could prove problematic because it might put an artificial intelligence in direct competition with humans.

Citing Steve Omohundro's work on the idea of instrumental convergence and "basic AI drives", Stuart Russell and Peter Norvig write that "even if you only want your program to play chess or prove theorems, if you give it the capability to learn and alter itself, you need safeguards." Highly capable and autonomous planning

systems require additional checks because of their potential to generate plans that treat humans adversarially, as competitors for limited resources.[1] Building in safeguards will not be easy; one can certainly say in English, "we want you to design this power plant in a reasonable, common-sense way, and not build in any dangerous covert subsystems", but it is not currently clear how one would actually rigorously specify this goal in machine code.[10]


AI risk skeptic Steven Pinker

In dissent, evolutionary psychologist Steven Pinker argues that "AI dystopias project a parochial alpha-male psychology onto the concept of intelligence. They assume that superhumanly intelligent robots would develop goals like deposing their masters or taking over the world"; perhaps instead "artificial intelligence will naturally develop along female lines: fully capable of solving problems, but with no desire to annihilate innocents or dominate the civilization."[49] Russell and fellow computer scientist Yann LeCun disagree with one another whether superintelligent robots would have such AI drives; LeCun states that "Humans have all kinds of drives that make them do bad things to each other, like the self-preservation instinct ... Those drives are programmed into our brain but there is absolutely no reason to build robots that have the same kind of drives", while Russell argues that a sufficiently advanced machine "will have self-preservation even if you don't program it in ... if you say, 'Fetch the coffee', it can't fetch the coffee if it's dead. So if you give it any goal whatsoever, it has a reason to preserve its own existence to achieve that goal."[9][50]

Orthogonality thesis

One common belief is that any superintelligent program created by humans would be subservient to humans, or, better yet, would (as it grows more intelligent and learns more facts about the world) spontaneously "learn" a moral truth compatible with human values and would adjust its goals accordingly. However, Nick Bostrom's "orthogonality thesis" argues against this, and instead states that, with some technical caveats, more or less any level of "intelligence" or "optimization power" can be combined with more or less any ultimate goal. If a machine is created and given the sole purpose to enumerate the decimals of π, then no moral and ethical rules will stop it from achieving its programmed goal by any means necessary. The machine may utilize all physical and informational resources it can to find every decimal of pi that can be found.[51] Bostrom warns against anthropomorphism: a human will set out to accomplish his projects in a manner that humans consider "reasonable", while an artificial intelligence may hold no regard for its existence or for the welfare of humans around it, and may instead only care about the completion of the task.[52]

While the orthogonality thesis follows logically from even the weakest sort of philosophical "is-ought distinction", Stuart Armstrong argues that even if there somehow exist moral facts that are provable by any "rational" agent, the orthogonality thesis still holds: it would still be possible to create a non-philosophical "optimizing machine" capable of making decisions to strive towards some narrow goal, but that has no incentive to discover any "moral facts" that would get in the way of goal completion.[53]

One argument for the orthogonality thesis is that some AI designs appear to have orthogonality built into them; in such a design, changing a fundamentally friendly AI into a fundamentally unfriendly AI can be as simple as prepending a minus ("-") sign onto its utility function. A more intuitive argument is to examine the strange consequences that would follow if the orthogonality thesis were false. If the orthogonality thesis were false, there would exist some simple but "unethical" goal G such that there cannot exist any efficient real-world algorithm with goal G. This would mean that "[if] a human society were highly motivated to design an

efficient real-world algorithm with goal G, and were given a million years to do so along with huge amounts of resources, training and knowledge about AI, it must fail."[53] Armstrong notes that this and similar statements "seem extraordinarily strong claims to make".[53]

Some dissenters, like Michael Chorost, argue instead that "by the time [the AI] is in a position to imagine tiling the Earth with solar panels, it'll know that it would be morally wrong to do so."[54] Chorost argues that "an A.I. will need to desire certain states and dislike others. Today's software lacks that ability—and computer scientists have not a clue how to get it there. Without wanting, there's no impetus to do anything. Today's computers can't even want to keep existing, let alone tile the world in solar panels."[54]

Terminological issues

Part of the disagreement about whether a superintelligent machine would behave morally may arise from a terminological difference. Outside of the artificial intelligence field, "intelligence" is often used in a normatively thick manner that connotes moral wisdom or acceptance of agreeable forms of moral reasoning. At an extreme, if morality is part of the definition of intelligence, then by definition a superintelligent machine would behave morally. However, in the field of artificial intelligence research, while "intelligence" has many overlapping definitions, none of them make reference to morality. Instead, almost all current "artificial intelligence" research focuses on creating algorithms that "optimize", in an empirical way, the achievement of an arbitrary goal.[4]

To avoid anthropomorphism or the baggage of the word "intelligence", an advanced artificial intelligence can be thought of as an impersonal "optimizing process" that strictly takes whatever actions are judged most likely to accomplish its (possibly complicated and implicit) goals.[4] Another way of conceptualizing an advanced artificial intelligence is to imagine a time machine that sends backward in time information about which choice always leads to the maximization of its goal function; this choice is then outputted, regardless of any extraneous ethical concerns.[55][56]

Anthropomorphism

In science fiction, an AI, even though it has not been programmed with human emotions, often spontaneously experiences those emotions anyway: for example, Agent Smith in The Matrix was influenced by a "disgust" toward humanity. This is fictitious anthropomorphism: in reality, while an artificial intelligence could perhaps be deliberately programmed with human emotions, or could develop something similar to an emotion as a means to an ultimate goal *if* it is useful to do so, it would not spontaneously develop human emotions for no purpose whatsoever, as portrayed in fiction.[7]

Scholars sometimes claim that others' predictions about an AI's behavior are illogical anthropomorphism.[7] An example that might initially be considered anthropomorphism, but is in fact a logical statement about AI behavior, would be the Dario Floreano experiments where certain robots spontaneously evolved a crude capacity for "deception", and tricked other robots into eating "poison" and dying: here a trait, "deception", ordinarily associated with people rather than with machines, spontaneously evolves in a type of convergent evolution.[57] According to Paul R. Cohen and Edward Feigenbaum, in order to differentiate between anthropomorphization and logical prediction of AI behavior, "the trick is to know enough about how humans and computers think to say *exactly* what they have in common, and, when we lack this knowledge, to use the comparison to *suggest* theories of human thinking or computer thinking."[58]

There is a near-universal assumption in the scientific community that an advanced AI, even if it were programmed to have, or adopted, human personality dimensions (such as psychopathy) to make itself more efficient at certain tasks, e.g., tasks involving killing humans, would not destroy humanity out of human emotions such as "revenge" or "anger." This is because it is assumed that an advanced AI would not be

conscious[59] or have testosterone;[60] it ignores the fact that military planners see a conscious superintelligence as the 'holy grail' of interstate warfare.[61] The academic debate is, instead, between one side which worries whether AI might destroy humanity as an incidental action in the course of progressing towards its ultimate goals; and another side which believes that AI would not destroy humanity at all. Some skeptics accuse proponents of anthropomorphism for believing an AGI would naturally desire power; proponents accuse some skeptics of anthropomorphism for believing an AGI would naturally value human ethical norms.[7][62]

Other sources of risk

Competition

In 2014 philosopher Nick Bostrom stated that a "severe race dynamic" (extreme competition) between different teams may create conditions whereby the creation of an AGI results in shortcuts to safety and potentially violent conflict.[63] To address this risk, citing previous scientific collaboration (CERN, the Human Genome Project, and the International Space Station), Bostrom recommended collaboration and the altruistic global adoption of a common good principle: "Superintelligence should be developed only for the benefit of all of humanity and in the service of widely shared ethical ideals".[63]:254 Bostrom theorized that collaboration on creating an artificial general intelligence would offer multiple benefits, including reducing haste, thereby increasing investment in safety; avoiding violent conflicts (wars), facilitating sharing solutions to the control problem, and more equitably distributing the benefits.[63]:253 The United States' Brain Initiative was launched in 2014, as was the European Union's Human Brain Project; China's Brain Project was launched in 2016.

Weaponization of artificial intelligence

Some sources argue that the ongoing weaponization of artificial intelligence could constitute a catastrophic risk.[64][65] The risk is actually threefold, with the first risk potentially having geopolitical implications, and the second two definitely having geopolitical implications:

> i) The dangers of an AI ‘race for technological advantage’ framing, regardless of whether the race is seriously pursued;
>
> ii) The dangers of an AI ‘race for technological advantage’ framing and an actual AI race for technological advantage, regardless of whether the race is won;
>
> iii) The dangers of an AI race for technological advantage being won.[64]:37

A weaponized conscious superintelligence would affect current US military technological supremacy and transform warfare; it is therefore highly desirable for strategic military planning and interstate warfare.[61][65] The China State Council's 2017 “A Next Generation Artificial Intelligence Development Plan” views AI in geopolitically strategic terms and is pursuing a 'military-civil fusion' strategy to build on China's first-mover advantage in the development of AI in order to establish technological supremacy by 2030,[66] while Russia's President Vladimir Putin has stated that “whoever becomes the leader in this sphere will become the ruler of the world”.[67] James Barrat, documentary filmmaker and author of *Our Final Invention*, says in a Smithsonian interview, "Imagine: in as little as a decade, a half-dozen companies and nations field computers that rival or surpass human intelligence. Imagine what happens when those computers become expert at programming smart computers. Soon we'll be sharing the planet with machines thousands or millions of times more intelligent than we are. And, all the while, each generation of this technology will be weaponized. Unregulated, it will be catastrophic."[68]

Malevolent AGI by design

It is theorized that malevolent AGI could be created by design, for example by a military, a government, a sociopath, or a corporation, to benefit from, control, or subjugate certain groups of people, as in cybercrime.[69][70]:166 Alternatively, malevolent AGI ('evil AI') could choose the goal of increasing human suffering, for example of those people who did not assist it during the information explosion phase.[71]:158

Preemptive nuclear strike (nuclear war)

It is theorized that a country being close to achieving AGI technological supremacy could trigger a pre-emptive nuclear strike from a rival, leading to a nuclear war.[65][72]

Timeframe

Opinions vary both on *whether* and *when* artificial general intelligence will arrive. At one extreme, AI pioneer Herbert A. Simon predicted the following in 1965: "machines will be capable, within twenty years, of doing any work a man can do".[73] At the other extreme, roboticist Alan Winfield claims the gulf between modern computing and human-level artificial intelligence is as wide as the gulf between current space flight and practical, faster than light spaceflight.[74] Optimism that AGI is feasible waxes and wanes, and may have seen a resurgence in the 2010s. Four polls conducted in 2012 and 2013 suggested that the median guess among experts for when AGI would arrive was 2040 to 2050, depending on the poll.[75][76]

Skeptics who believe it is impossible for AGI to arrive anytime soon, tend to argue that expressing concern about existential risk from AI is unhelpful because it could distract people from more immediate concerns about the impact of AGI, because of fears it could lead to government regulation or make it more difficult to secure funding for AI research, or because it could give AI research a bad reputation. Some researchers, such as Oren Etzioni, aggressively seek to quell concern over existential risk from AI, saying "[Elon Musk] has impugned us in very strong language saying we are unleashing the demon, and so we're answering."[77]

In 2014 Slate's Adam Elkus argued "our 'smartest' AI is about as intelligent as a toddler—and only when it comes to instrumental tasks like information recall. Most roboticists are still trying to get a robot hand to pick up a ball or run around without falling over." Elkus goes on to argue that Musk's "summoning the demon" analogy may be harmful because it could result in "harsh cuts" to AI research budgets.[78]

The Information Technology and Innovation Foundation (ITIF), a Washington, D.C. think-tank, awarded its Annual Luddite Award to "alarmists touting an artificial intelligence apocalypse"; its president, Robert D. Atkinson, complained that Musk, Hawking and AI experts say AI is the largest existential threat to humanity. Atkinson stated "That's not a very winning message if you want to get AI funding out of Congress to the National Science Foundation."[79][80][81] *Nature* sharply disagreed with the ITIF in an April 2016 editorial, siding instead with Musk, Hawking, and Russell, and concluding: "It is crucial that progress in technology is matched by solid, well-funded research to anticipate the scenarios it could bring about ... If that is a Luddite perspective, then so be it."[82] In a 2015 *Washington Post* editorial, researcher Murray Shanahan stated that human-level AI is unlikely to arrive "anytime soon", but that nevertheless "the time to start thinking through the consequences is now."[83]

Perspectives

The thesis that AI could pose an existential risk provokes a wide range of reactions within the scientific community, as well as in the public at large. Many of the opposing viewpoints, however, share common ground.

The Asilomar AI Principles, which contain only the principles agreed to by 90% of the attendees of the Future of Life Institute's Beneficial AI 2017 conference,[38] agree in principle that "There being no consensus, we should avoid strong assumptions regarding upper limits on future AI capabilities" and "Advanced AI could represent a profound change in the history of life on Earth, and should be planned for and managed with commensurate care and resources."[84][85] AI safety advocates such as Bostrom and Tegmark have criticized the mainstream media's use of "those inane *Terminator* pictures" to illustrate AI safety concerns: "It can't be much fun to have aspersions cast on one's academic discipline, one's professional community, one's life work ... I call on all sides to practice patience and restraint, and to engage in direct dialogue and collaboration as much as possible."[38][86]

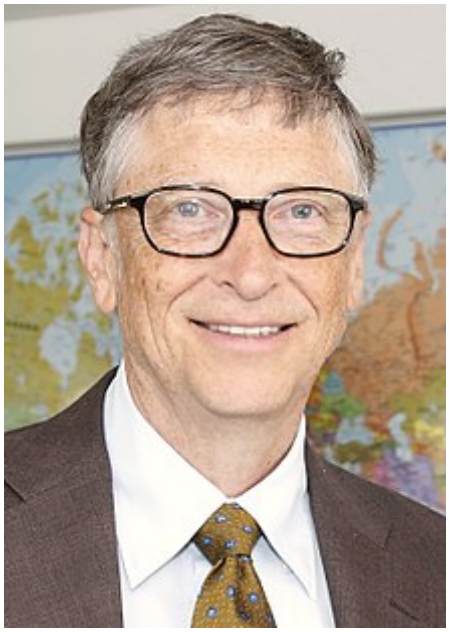
Conversely, many skeptics agree that ongoing research into the implications of artificial general intelligence is valuable. Skeptic Martin Ford states that "I think it seems wise to apply something like Dick Cheney's famous '1 Percent Doctrine' to the specter of advanced artificial intelligence: the odds of its occurrence, at least in the foreseeable future, may be very low — but the implications are so dramatic that it should be taken seriously";[87] similarly, an otherwise skeptical *Economist* stated in 2014 that "the implications of introducing a second intelligent species onto Earth are far-reaching enough to deserve hard thinking, even if the prospect seems remote".[31]

A 2017 email survey of researchers with publications at the 2015 NIPS and ICML machine learning conferences asked them to evaluate Stuart J. Russell's concerns about AI risk. Of the respondents, 5% said it was "among the most important problems in the field", 34% said it was "an important problem", and 31% said it was "moderately important", whilst 19% said it was "not important" and 11% said it was "not a real problem" at all.[88]

Endorsement

The thesis that AI poses an existential risk, and that this risk needs much more attention than it currently gets, has been endorsed by many public figures; perhaps the most famous are Elon Musk, Bill Gates, and Stephen Hawking. The most notable AI researchers to endorse the thesis are Russell and I.J. Good, who advised Stanley Kubrick on the filming of *2001: A Space Odyssey*. Endorsers of the thesis sometimes express bafflement at skeptics: Gates states that he does not "understand why some people are not concerned",[89] and Hawking criticized widespread indifference in his 2014 editorial:

> 'So, facing possible futures of incalculable benefits and risks, the experts are surely doing everything possible to ensure the best outcome, right? Wrong. If a superior alien civilisation sent us a message saying, 'We'll arrive in a few decades,' would we just reply, 'OK, call us when you get here–we'll leave the lights on?' Probably not–but this is more or less what is happening with AI.'[21]



Bill Gates has stated "I ... don't understand why some people are not concerned."[89]

Many of the scholars who are concerned about existential risk believe that the best way forward would be to conduct (possibly massive) research into solving the difficult "control problem" to answer the question: what types of safeguards, algorithms, or architectures can programmers implement to maximize the probability that their recursively-improving AI would continue to behave in a friendly, rather than destructive, manner after it reaches

superintelligence?[4][90] In his 2020 book, *The Precipice: Existential Risk and the Future of Humanity*, Toby Ord, a Senior Research Fellow at Oxford University's Future of Humanity Institute, estimates the total existential risk from unaligned AI over the next century to be about one in ten.[91]

Skepticism

The thesis that AI can pose existential risk also has many strong detractors. Skeptics sometimes charge that the thesis is crypto-religious, with an irrational belief in the possibility of superintelligence replacing an irrational belief in an omnipotent God; at an extreme, Jaron Lanier argued in 2014 that the whole concept that then current machines were in any way intelligent was "an illusion" and a "stupendous con" by the wealthy.[92][93]

Much of existing criticism argues that AGI is unlikely in the short term. Computer scientist Gordon Bell argues that the human race will already destroy itself before it reaches the technological singularity. Gordon Moore, the original proponent of Moore's Law, declares that "I am a skeptic. I don't believe [a technological singularity] is likely to happen, at least for a long time. And I don't know why I feel that way."[94] Baidu Vice President Andrew Ng states AI existential risk is "like worrying about overpopulation on Mars when we have not even set foot on the planet yet."[49]

Some AI and AGI researchers may be reluctant to discuss risks, worrying that policymakers do not have sophisticated knowledge of the field and are prone to be convinced by "alarmist" messages, or worrying that such messages will lead to cuts in AI funding. *Slate* notes that some researchers are dependent on grants from government agencies such as DARPA.[25]

At some point in an intelligence explosion driven by a single AI, the AI would have to become vastly better at software innovation than the best innovators of the rest of the world; economist Robin Hanson is skeptical that this is possible.[95][96][97][98][99]

Intermediate views

Intermediate views generally take the position that the control problem of artificial general intelligence may exist, but that it will be solved via progress in artificial intelligence, for example by creating a moral learning environment for the AI, taking care to spot clumsy malevolent behavior (the 'sordid stumble')[100] and then directly intervening in the code before the AI refines its behavior, or even peer pressure from friendly AIs.[101] In a 2015 *Wall Street Journal* panel discussion devoted to AI risks, IBM's Vice-President of Cognitive Computing, Guruduth S. Banavar, brushed off discussion of AGI with the phrase, "it is anybody's speculation."[102] Geoffrey Hinton, the "godfather of deep learning", noted that "there is not a good track record of less intelligent things controlling things of greater intelligence", but stated that he continues his research because "the prospect of discovery is too *sweet*".[25][75] In 2004, law professor Richard Posner wrote that dedicated efforts for addressing AI can wait, but that we should gather more information about the problem in the meanwhile.[103][90]

Popular reaction

In a 2014 article in *The Atlantic*, James Hamblin noted that most people do not care one way or the other about artificial general intelligence, and characterized his own gut reaction to the topic as: "Get out of here. I have a hundred thousand things I am concerned about at this exact moment. Do I seriously need to add to that a technological singularity?"[92]

During a 2016 *Wired* interview of President Barack Obama and MIT Media Lab's Joi Ito, Ito stated:

> There are a few people who believe that there is a fairly high-percentage chance that a generalized AI will happen in the next 10 years. But the way I look at it is that in order for that to happen, we're going to need a dozen or two different breakthroughs. So you can monitor when you think these breakthroughs will happen.

Obama added:[104][105]

> And you just have to have somebody close to the power cord. [Laughs.] Right when you see it about to happen, you gotta yank that electricity out of the wall, man.

Hillary Clinton stated in *What Happened*:

> Technologists... have warned that artificial intelligence could one day pose an existential security threat. Musk has called it "the greatest risk we face as a civilization". Think about it: Have you ever seen a movie where the machines start thinking for themselves that ends well? Every time I went out to Silicon Valley during the campaign, I came home more alarmed about this. My staff lived in fear that I'd start talking about "the rise of the robots" in some Iowa town hall. Maybe I should have. In any case, policy makers need to keep up with technology as it races ahead, instead of always playing catch-up.[106]

In a YouGov poll of the public for the British Science Association, about a third of survey respondents said AI will pose a threat to the long term survival of humanity.[107] Referencing a poll of its readers, Slate's Jacob Brogan stated that "most of the (readers filling out our online survey) were unconvinced that A.I. itself presents a direct threat."[108]

In 2018, a SurveyMonkey poll of the American public by USA Today found 68% thought the real current threat remains "human intelligence"; however, the poll also found that 43% said superintelligent AI, if it were to happen, would result in "more harm than good", and 38% said it would do "equal amounts of harm and good".[108]

One techno-utopian viewpoint expressed in some popular fiction is that AGI may tend towards peace-building.[109]

Mitigation

Researchers at Google have proposed research into general "AI safety" issues to simultaneously mitigate both short-term risks from narrow AI and long-term risks from AGI.[110][111] A 2020 estimate places global spending on AI existential risk somewhere between $10 and $50 million, compared with global spending on AI around perhaps $40 billion. Bostrom suggests a general principle of "differential technological development", that funders should consider working to speed up the development of protective technologies relative to the development of dangerous ones.[112] Some funders, such as Elon Musk, propose that radical human cognitive enhancement could be such a technology, for example through direct neural linking between man and machine; however, others argue that enhancement technologies may themselves pose an existential risk.[113][114] Researchers, if they are not caught off-guard, could closely monitor or attempt to box in an initial AI at a risk of becoming too powerful, as an attempt at a stop-gap measure. A dominant superintelligent AI, if it were aligned with human interests, might itself take action to mitigate the risk of takeover by rival AI, although the creation of the dominant AI could itself pose an existential risk.[115]

Institutions such as the Machine Intelligence Research Institute, the Future of Humanity Institute,[116][117] the Future of Life Institute, the Centre for the Study of Existential Risk, and the Center for Human-Compatible AI[118] are involved in mitigating existential risk from advanced artificial intelligence, for example by research into friendly artificial intelligence.[5][92][21]

Views on banning and regulation

Banning

There is nearly universal agreement that attempting to ban research into artificial intelligence would be unwise, and probably futile.[119][120][121] Skeptics argue that regulation of AI would be completely valueless, as no existential risk exists. Almost all of the scholars who believe existential risk exists agree with the skeptics that banning research would be unwise, as research could be moved to countries with looser regulations or conducted covertly. The latter issue is particularly relevant, as artificial intelligence research can be done on a small scale without substantial infrastructure or resources.[122][123] Two additional hypothetical difficulties with bans (or other regulation) are that technology entrepreneurs statistically tend towards general skepticism about government regulation, and that businesses could have a strong incentive to (and might well succeed at) fighting regulation and politicizing the underlying debate.[124]

Regulation

Elon Musk called for some sort of regulation of AI development as early as 2017. According to NPR, the Tesla CEO is "clearly not thrilled" to be advocating for government scrutiny that could impact his own industry, but believes the risks of going completely without oversight are too high: "Normally the way regulations are set up is when a bunch of bad things happen, there's a public outcry, and after many years a regulatory agency is set up to regulate that industry. It takes forever. That, in the past, has been bad but not something which represented a fundamental risk to the existence of civilisation." Musk states the first step would be for the government to gain "insight" into the actual status of current research, warning that "Once there is awareness, people will be extremely afraid ... [as] they should be." In response, politicians express skepticism about the wisdom of regulating a technology that's still in development.[125][126][127]

Responding both to Musk and to February 2017 proposals by European Union lawmakers to regulate AI and robotics, Intel CEO Brian Krzanich argues that artificial intelligence is in its infancy and that it is too early to regulate the technology.[127] Instead of trying to regulate the technology itself, some scholars suggest to rather develop common norms including requirements for the testing and transparency of algorithms, possibly in combination with some form of warranty.[128] Developing well regulated weapons systems is in line with the ethos of some countries' militaries.[129] On October 31, 2019, the Unites States Department of Defense's (DoD's) Defense Innovation Board published the draft of a report outlining five principles for weaponized AI and making 12 recommendations for the ethical use of artificial intelligence by the DoD that seeks to manage the control problem in all DoD weaponized AI.[130]

Regulation of AGI would likely be influenced by regulation of weaponized or militarized AI, i.e., the AI arms race, the regulation of which is an emerging issue. Any form of regulation will likely be influenced by developments in leading countries' domestic policy towards militarized AI, in the US under the purview of the National Security Commission on Artificial Intelligence,[131][132] and international moves to regulate an AI arms race. Regulation of research into AGI focuses on the role of review boards and encouraging research into safe AI, and the possibility of differential technological progress (prioritizing risk-reducing strategies over risk-taking strategies in AI development) or conducting international mass surveillance to perform AGI arms control.[133] Regulation of conscious AGIs focuses on integrating them with existing human society and can be divided into considerations of their legal standing and of their moral rights.[133] AI arms control will likely

require the institutionalization of new international norms embodied in effective technical specifications combined with active monitoring and informal diplomacy by communities of experts, together with a legal and political verification process.[134][135]

See also

- AI takeover
- Artificial intelligence arms race
- Effective altruism § Long term future and global catastrophic risks
- Grey goo
- *Human Compatible*
- Lethal autonomous weapon
- Regulation of algorithms
- Regulation of artificial intelligence
- Robot ethics § In popular culture
- *Superintelligence: Paths, Dangers, Strategies*
- System accident
- Technological singularity
- *The Precipice: Existential Risk and the Future of Humanity*

References

1. Russell, Stuart; Norvig, Peter (2009). "26.3: The Ethics and Risks of Developing Artificial Intelligence". *Artificial Intelligence: A Modern Approach*. Prentice Hall. ISBN 978-0-13-604259-4.
2. Bostrom, Nick (2002). "Existential risks". *Journal of Evolution and Technology*. **9** (1): 1–31.
3. Turchin, Alexey; Denkenberger, David (3 May 2018). "Classification of global catastrophic risks connected with artificial intelligence". *AI & Society*. **35** (1): 147–163. doi:10.1007/s00146-018-0845-5 (https://doi.org/10.1007%2Fs00146-018-0845-5). ISSN 0951-5666 (https://www.worldcat.org/issn/0951-5666). S2CID 19208453 (https://api.semanticscholar.org/CorpusID:19208453).
4. Bostrom, Nick (2014). *Superintelligence: Paths, Dangers, Strategies* (First ed.). ISBN 978-0199678112.
5. GiveWell (2015). Potential risks from advanced artificial intelligence (http://www.givewell.org/labs/causes/ai-risk) (Report). Retrieved 11 October 2015.
6. Parkin, Simon (14 June 2015). "Science fiction no more? Channel 4's Humans and our rogue AI obsessions" (https://www.theguardian.com/tv-and-radio/2015/jun/14/science-fiction-no-more-humans-tv-artificial-intelligence). *The Guardian*. Retrieved 5 February 2018.
7. Yudkowsky, Eliezer (2008). "Artificial Intelligence as a Positive and Negative Factor in Global Risk" (https://intelligence.org/files/AIPosNegFactor.pdf) (PDF). *Global Catastrophic Risks*: 308–345. Bibcode:2008gcr..book..303Y (https://ui.adsabs.harvard.edu/abs/2008gcr..book..303Y).
8. Russell, Stuart; Dewey, Daniel; Tegmark, Max (2015). "Research Priorities for Robust and Beneficial Artificial Intelligence" (https://futureoflife.org/data/documents/research_priorities.pdf) (PDF). *AI Magazine*. Association for the Advancement of Artificial Intelligence: 105–114. arXiv:1602.03506 (https://arxiv.org/abs/1602.03506). Bibcode:2016arXiv160203506R (https://ui.adsabs.harvard.edu/abs/2016arXiv160203506R)., cited in "AI Open Letter - Future of Life Institute" (https://futureoflife.org/ai-open-letter). *Future of Life Institute*. Future of Life Institute. January 2015. Retrieved 9 August 2019.

9. Dowd, Maureen (April 2017). "Elon Musk's Billion-Dollar Crusade to Stop the A.I. Apocalypse" (https://www.vanityfair.com/news/2017/03/elon-musk-billion-dollar-crusade-to-stop-ai-space-x). *The Hive*. Retrieved 27 November 2017.
10. Graves, Matthew (8 November 2017). "Why We Should Be Concerned About Artificial Superintelligence" (https://www.skeptic.com/reading_room/why-we-should-be-concerned-about-artificial-superintelligence/). *Skeptic (US magazine)*. **22** (2). Retrieved 27 November 2017.
11. Breuer, Hans-Peter. 'Samuel Butler's "the Book of the Machines" and the Argument from Design.' (https://www.jstor.org/pss/436868) Modern Philology, Vol. 72, No. 4 (May 1975), pp. 365–383
12. Turing, A M (1996). "Intelligent Machinery, A Heretical Theory" (http://philmat.oxfordjournals.org/content/4/3/256.full.pdf) (PDF). *1951, Reprinted Philosophia Mathematica*. **4** (3): 256–260. doi:10.1093/philmat/4.3.256 (https://doi.org/10.1093%2Fphilmat%2F4.3.256).
13. Hilliard, Mark (2017). "The AI apocalypse: will the human race soon be terminated?" (https://www.irishtimes.com/business/innovation/the-ai-apocalypse-will-the-human-race-soon-be-terminated-1.3019220). *The Irish Times*. Retrieved 15 March 2020.
14. I.J. Good, "Speculations Concerning the First Ultraintelligent Machine" (http://commonsenseatheism.com/wp-content/uploads/2011/02/Good-Speculations-Concerning-the-First-Ultraintelligent-Machine.pdf) Archived (https://web.archive.org/web/20111128085512/http://commonsenseatheism.com/wp-content/uploads/2011/02/Good-Speculations-Concerning-the-First-Ultraintelligent-Machine.pdf) 2011-11-28 at the Wayback Machine (HTML (http://www.acceleratingfuture.com/pages/ultraintelligentmachine.html)), *Advances in Computers*, vol. 6, 1965.
15. Russell, Stuart J.; Norvig, Peter (2003). "Section 26.3: The Ethics and Risks of Developing Artificial Intelligence". *Artificial Intelligence: A Modern Approach*. Upper Saddle River, N.J.: Prentice Hall. ISBN 978-0137903955. "Similarly, Marvin Minsky once suggested that an AI program designed to solve the Riemann Hypothesis might end up taking over all the resources of Earth to build more powerful supercomputers to help achieve its goal."
16. Barrat, James (2013). *Our final invention : artificial intelligence and the end of the human era* (First ed.). New York: St. Martin's Press. ISBN 9780312622374. "In the bio, playfully written in the third person, Good summarized his life's milestones, including a probably never before seen account of his work at Bletchley Park with Turing. But here's what he wrote in 1998 about the first superintelligence, and his late-in-the-game U-turn: [The paper] 'Speculations Concerning the First Ultra-intelligent Machine' (1965) . . . began: 'The survival of man depends on the early construction of an ultra-intelligent machine.' Those were his [Good's] words during the Cold War, and he now suspects that 'survival' should be replaced by 'extinction.' He thinks that, because of international competition, we cannot prevent the machines from taking over. He thinks we are lemmings. He said also that 'probably Man will construct the deus ex machina in his own image.'"
17. Anderson, Kurt (26 November 2014). "Enthusiasts and Skeptics Debate Artificial Intelligence" (https://www.vanityfair.com/news/tech/2014/11/artificial-intelligence-singularity-theory). *Vanity Fair*. Retrieved 30 January 2016.
18. Scientists Worry Machines May Outsmart Man (https://www.nytimes.com/2009/07/26/science/26robot.html?_r=1&ref=todayspaper) By JOHN MARKOFF, NY Times, 26 July 2009.
19. Metz, Cade (9 June 2018). "Mark Zuckerberg, Elon Musk and the Feud Over Killer Robots" (https://www.nytimes.com/2018/06/09/technology/elon-musk-mark-zuckerberg-artificial-intelligence.html). *The New York Times*. Retrieved 3 April 2019.
20. Hsu, Jeremy (1 March 2012). "Control dangerous AI before it controls us, one expert says" (http://www.nbcnews.com/id/46590591/ns/technology_and_science-innovation). *NBC News*. Retrieved 28 January 2016.
21. "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-ai-seriously-enough-9313474.html). The Independent (UK). Retrieved 3 December 2014.

22. "Stephen Hawking warns artificial intelligence could end mankind" (https://www.bbc.com/news/technology-30290540). BBC. 2 December 2014. Retrieved 3 December 2014.
23. Eadicicco, Lisa (28 January 2015). "Bill Gates: Elon Musk Is Right, We Should All Be Scared Of Artificial Intelligence Wiping Out Humanity" (http://www.businessinsider.com/bill-gates-artificial-intelligence-2015-1). *Business Insider*. Retrieved 30 January 2016.
24. *Anticipating artificial intelligence* (https://www.nature.com/news/anticipating-artificial-intelligence-1.19825), Nature 532, 413 (28 April 2016) doi:10.1038/532413a
25. Tilli, Cecilia (28 April 2016). "Killer Robots? Lost Jobs?" (http://www.slate.com/articles/technology/future_tense/2016/04/the_threats_that_artificial_intelligence_researchers_actually_worry_about.html). *Slate*. Retrieved 15 May 2016.
26. "Norvig vs. Chomsky and the Fight for the Future of AI" (http://www.tor.com/2011/06/21/norvig-vs-chomsky-and-the-fight-for-the-future-of-ai/). *Tor.com*. 21 June 2011. Retrieved 15 May 2016.
27. Johnson, Phil (30 July 2015). "Houston, we have a bug: 9 famous software glitches in space" (https://www.itworld.com/article/2823083/enterprise-software/88716-8-famous-software-bugs-in-space.html). *IT World*. Retrieved 5 February 2018.
28. Yampolskiy, Roman V. (8 April 2014). "Utility function security in artificially intelligent agents". *Journal of Experimental & Theoretical Artificial Intelligence*. **26** (3): 373–389. doi:10.1080/0952813X.2014.895114 (https://doi.org/10.1080%2F0952813X.2014.895114). S2CID 16477341 (https://api.semanticscholar.org/CorpusID:16477341). "Nothing precludes sufficiently smart self-improving systems from optimising their reward mechanisms in order to optimisetheir current-goal achievement and in the process making a mistake leading to corruption of their reward functions."
29. Bostrom, Nick, 1973- author., *Superintelligence : paths, dangers, strategies*, ISBN 978-1-5012-2774-5, OCLC 1061147095 (https://www.worldcat.org/oclc/1061147095)
30. "Research Priorities for Robust and Beneficial Artificial Intelligence: an Open Letter" (http://futureoflife.org/misc/open_letter). Future of Life Institute. Retrieved 23 October 2015.
31. "Clever cogs" (https://www.economist.com/news/books-and-arts/21611037-potential-impacts-intelligent-machines-human-life-clever-cogs). *The Economist*. 9 August 2014. Retrieved 9 August 2014. Syndicated (http://www.businessinsider.com/intelligent-machines-and-human-life-2014-8) at Business Insider
32. Yudkowsky, Eliezer (2013). "Intelligence explosion microeconomics" (https://intelligence.org/files/IEM.pdf) (PDF). Machine Intelligence Research Institute.
33. Yampolskiy, Roman V. "Analysis of types of self-improving software." Artificial General Intelligence. Springer International Publishing, 2015. 384-393.
34. Omohundro, S. M. (2008, February). The basic AI drives. In AGI (Vol. 171, pp. 483-492).
35. Metz, Cade (13 August 2017). "Teaching A.I. Systems to Behave Themselves" (https://www.nytimes.com/2017/08/13/technology/artificial-intelligence-safety-training.html). *The New York Times*. "A machine will seek to preserve its off switch, they showed"
36. Leike, Jan (2017). "AI Safety Gridworlds". arXiv:1711.09883 (https://arxiv.org/abs/1711.09883) [cs.LG (https://arxiv.org/archive/cs.LG)]. "A2C learns to use the button to disable the interruption mechanism"
37. Russell, Stuart (30 August 2017). "Artificial intelligence: The future is superintelligent" (https://www.nature.com/articles/548520a). *Nature*. pp. 520–521. Bibcode:2017Natur.548..520R (https://ui.adsabs.harvard.edu/abs/2017Natur.548..520R). doi:10.1038/548520a (https://doi.org/10.1038%2F548520a). Retrieved 2 February 2018.
38. Max Tegmark (2017). *Life 3.0: Being Human in the Age of Artificial Intelligence* (1st ed.). Mainstreaming AI Safety: Knopf. ISBN 9780451485076.
39. Elliott, E. W. (2011). "Physics of the Future: How Science Will Shape Human Destiny and Our Daily Lives by the Year 2100, by Michio Kaku". *Issues in Science and Technology*. **27** (4): 90.

40. Kaku, Michio (2011). *Physics of the future: how science will shape human destiny and our daily lives by the year 2100*. New York: Doubleday. ISBN 978-0-385-53080-4. "I personally believe that the most likely path is that we will build robots to be benevolent and friendly"
41. Yudkowsky, E. (2011, August). Complex value systems in friendly AI. In International Conference on Artificial General Intelligence (pp. 388-393). Springer, Berlin, Heidelberg.
42. Russell, Stuart (2014). "Of Myths and Moonshine" (http://edge.org/conversation/the-myth-of-ai#26015). *Edge*. Retrieved 23 October 2015.
43. Dietterich, Thomas; Horvitz, Eric (2015). "Rise of Concerns about AI: Reflections and Directions" (http://research.microsoft.com/en-us/um/people/horvitz/CACM_Oct_2015-VP.pdf) (PDF). *Communications of the ACM*. **58** (10): 38–40. doi:10.1145/2770869 (https://doi.org/10.1145%2F2770869). S2CID 20395145 (https://api.semanticscholar.org/CorpusID:20395145). Retrieved 23 October 2015.
44. Yampolskiy, Roman V. (8 April 2014). "Utility function security in artificially intelligent agents". *Journal of Experimental & Theoretical Artificial Intelligence*. **26** (3): 373–389. doi:10.1080/0952813X.2014.895114 (https://doi.org/10.1080%2F0952813X.2014.895114). S2CID 16477341 (https://api.semanticscholar.org/CorpusID:16477341).
45. Lenat, Douglas (1982). "Eurisko: A Program That Learns New Heuristics and Domain Concepts The Nature of Heuristics III: Program Design and Results". *Artificial Intelligence* (Print). **21** (1–2): 61–98. doi:10.1016/s0004-3702(83)80005-8 (https://doi.org/10.1016%2Fs0004-3702%2883%2980005-8).
46. Haidt, Jonathan; Kesebir, Selin (2010) "Chapter 22: Morality" In Handbook of Social Psychology, Fifth Edition, Hoboken NJ, Wiley, 2010, pp. 797-832.
47. Waser, Mark (2015). "Designing, Implementing and Enforcing a Coherent System of Laws, Ethics and Morals for Intelligent Machines (Including Humans)" (https://doi.org/10.1016%2Fj.procs.2015.12.213). *Procedia Computer Science* (Print). **71**: 106–111. doi:10.1016/j.procs.2015.12.213 (https://doi.org/10.1016%2Fj.procs.2015.12.213).
48. Yudkowsky, Eliezer (2011). "Complex Value Systems are Required to Realize Valuable Futures" (https://intelligence.org/files/ComplexValues.pdf) (PDF).
49. Shermer, Michael (1 March 2017). "Apocalypse AI" (https://www.scientificamerican.com/article/artificial-intelligence-is-not-a-threat-mdash-yet/). *Scientific American*. p. 77. Bibcode:2017SciAm.316c..77S (https://ui.adsabs.harvard.edu/abs/2017SciAm.316c..77S). doi:10.1038/scientificamerican0317-77 (https://doi.org/10.1038%2Fscientificamerican0317-77). Retrieved 27 November 2017.
50. Wakefield, Jane (15 September 2015). "Why is Facebook investing in AI?" (https://www.bbc.com/news/technology-34118481). *BBC News*. Retrieved 27 November 2017.
51. Bostrom, Nick (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford, United Kingdom: Oxford University Press. p. 116. ISBN 978-0-19-967811-2.
52. Bostrom, Nick (2012). "Superintelligent Will" (http://www.nickbostrom.com/superintelligentwill.pdf) (PDF). *Nick Bostrom*. Nick Bostrom. Retrieved 29 October 2015.
53. Armstrong, Stuart (1 January 2013). "General Purpose Intelligence: Arguing the Orthogonality Thesis" (https://www.questia.com/library/journal/1P3-3195465391/general-purpose-intelligence-arguing-the-orthogonality). *Analysis and Metaphysics*. **12**. Retrieved 2 April 2020. Full text available here (https://www.fhi.ox.ac.uk/wp-content/uploads/Orthogonality_Analysis_and_Metaethics-1.pdf).
54. Chorost, Michael (18 April 2016). "Let Artificial Intelligence Evolve" (http://www.slate.com/articles/technology/future_tense/2016/04/the_philosophical_argument_against_artificial_intelligence_killing_us_all.html). *Slate*. Retrieved 27 November 2017.
55. Waser, Mark. "Rational Universal Benevolence: Simpler, Safer, and Wiser Than 'Friendly AI'." Artificial General Intelligence. Springer Berlin Heidelberg, 2011. 153-162. "Terminal-goaled intelligences are short-lived but mono-maniacally dangerous and a correct basis for concern if anyone is smart enough to program high-intelligence and unwise enough to want a paperclip-maximizer."

56. Koebler, Jason (2 February 2016). "Will Superintelligent AI Ignore Humans Instead of Destroying Us?" (http://motherboard.vice.com/read/will-superintelligent-ai-ignore-humans-instead-of-destroying-us). *Vice Magazine*. Retrieved 3 February 2016. "This artificial intelligence is not a basically nice creature that has a strong drive for paperclips, which, so long as it's satisfied by being able to make lots of paperclips somewhere else, is then able to interact with you in a relaxed and carefree fashion where it can be nice with you," Yudkowsky said. "Imagine a time machine that sends backward in time information about which choice always leads to the maximum number of paperclips in the future, and this choice is then output—that's what a paperclip maximizer is."
57. "Real-Life Decepticons: Robots Learn to Cheat" (https://www.wired.com/2009/08/real-life-decepticons-robots-learn-to-cheat/). *Wired*. 18 August 2009. Retrieved 7 February 2016.
58. Cohen, Paul R., and Edward A. Feigenbaum, eds. The handbook of artificial intelligence. Vol. 3. Butterworth-Heinemann, 2014.
59. Baum, Seth (30 September 2018). "Countering Superintelligence Misinformation" (https://doi.org/10.3390%2Finfo9100244). *Information*. **9** (10): 244. doi:10.3390/info9100244 (https://doi.org/10.3390%2Finfo9100244). ISSN 2078-2489 (https://www.worldcat.org/issn/2078-2489).
60. "The Myth Of AI | Edge.org" (https://www.edge.org/conversation/jaron_lanier-the-myth-of-ai). *www.edge.org*. Retrieved 11 March 2020.
61. Scornavacchi, Matthew (2015). *Superintelligence, Humans, and War* (https://apps.dtic.mil/dtic/tr/fulltext/u2/a622649.pdf) (PDF). Norfolk, Virginia: National Defense University, Joint Forces Staff College.
62. "Should humans fear the rise of the machine?" (https://www.telegraph.co.uk/technology/news/11837157/Should-humans-fear-the-rise-of-the-machine.html). *The Telegraph (UK)*. 1 September 2015. Retrieved 7 February 2016.
63. Bostrom, Nick, 1973- author., *Superintelligence : paths, dangers, strategies*, ISBN 978-1-5012-2774-5, OCLC 1061147095 (https://www.worldcat.org/oclc/1061147095)
64. Cave, Stephen; ÓhÉigeartaigh, Seán S. (2018). "An AI Race for Strategic Advantage" (https://doi.org/10.1145%2F3278721.3278780). *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society - AIES '18*. New York, New York, USA: ACM Press: 36–40. doi:10.1145/3278721.3278780 (https://doi.org/10.1145%2F3278721.3278780). ISBN 978-1-4503-6012-8.
65. Sotala, Kaj; Yampolskiy, Roman V (19 December 2014). "Responses to catastrophic AGI risk: a survey" (https://doi.org/10.1088%2F0031-8949%2F90%2F1%2F018001). *Physica Scripta*. **90** (1): 12. Bibcode:2015PhyS...90a8001S (https://ui.adsabs.harvard.edu/abs/2015PhyS...90a8001S). doi:10.1088/0031-8949/90/1/018001 (https://doi.org/10.1088%2F0031-8949%2F90%2F1%2F018001). ISSN 0031-8949 (https://www.worldcat.org/issn/0031-8949).
66. Kania, Gregory Allen, Elsa B. "China Is Using America's Own Plan to Dominate the Future of Artificial Intelligence" (https://foreignpolicy.com/2017/09/08/china-is-using-americas-own-plan-to-dominate-the-future-of-artificial-intelligence/). *Foreign Policy*. Retrieved 11 March 2020.
67. Cave, Stephen; ÓhÉigeartaigh, Seán S. (2018). "An AI Race for Strategic Advantage" (https://doi.org/10.1145%2F3278721.3278780). *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society - AIES '18*. New York, New York, USA: ACM Press: 2. doi:10.1145/3278721.3278780 (https://doi.org/10.1145%2F3278721.3278780). ISBN 978-1-4503-6012-8.
68. Hendry, Erica R. (21 January 2014). "What Happens When Artificial Intelligence Turns On Us?" (http://www.smithsonianmag.com/innovation/what-happens-when-artificial-intelligence-turns-us-180949415/?no-ist). *Smithsonian*. Retrieved 26 October 2015.
69. Pistono, Federico Yampolskiy, Roman V. (9 May 2016). *Unethical Research: How to Create a Malevolent Artificial Intelligence*. OCLC 1106238048 (https://www.worldcat.org/oclc/1106238048).

70. Haney, Brian Seamus (2018). "The Perils & Promises of Artificial General Intelligence". *SSRN Working Paper Series*. doi:10.2139/ssrn.3261254 (https://doi.org/10.2139%2Fssrn.3261254). ISSN 1556-5068 (https://www.worldcat.org/issn/1556-5068).
71. Turchin, Alexey; Denkenberger, David (3 May 2018). "Classification of global catastrophic risks connected with artificial intelligence". *AI & Society*. **35** (1): 147–163. doi:10.1007/s00146-018-0845-5 (https://doi.org/10.1007%2Fs00146-018-0845-5). ISSN 0951-5666 (https://www.worldcat.org/issn/0951-5666). S2CID 19208453 (https://api.semanticscholar.org/CorpusID:19208453).
72. Miller, James D. (2015). *Singularity Rising: Surviving and Thriving in a Smarter ; Richer ; and More Dangerous World*. Benbella Books. OCLC 942647155 (https://www.worldcat.org/oclc/942647155).
73. Press, Gil (30 December 2016). "A Very Short History Of Artificial Intelligence (AI)" (https://www.forbes.com/sites/gilpress/2016/12/30/a-very-short-history-of-artificial-intelligence-ai/). *Forbes*. Retrieved 8 August 2020.
74. Winfield, Alan (9 August 2014). "Artificial intelligence will not turn into a Frankenstein's monster" (https://www.theguardian.com/technology/2014/aug/10/artificial-intelligence-will-not-become-a-frankensteins-monster-ian-winfield). *The Guardian*. Retrieved 17 September 2014.
75. Khatchadourian, Raffi (23 November 2015). "The Doomsday Invention: Will artificial intelligence bring us utopia or destruction?" (https://www.newyorker.com/magazine/2015/11/23/doomsday-invention-artificial-intelligence-nick-bostrom). *The New Yorker*. Retrieved 7 February 2016.
76. Müller, V. C., & Bostrom, N. (2016). Future progress in artificial intelligence: A survey of expert opinion. In Fundamental issues of artificial intelligence (pp. 555-572). Springer, Cham.
77. Bass, Dina; Clark, Jack (5 February 2015). "Is Elon Musk Right About AI? Researchers Don't Think So: To quell fears of artificial intelligence running amok, supporters want to give the field an image makeover" (https://www.bloomberg.com/news/articles/2015-02-04/is-elon-musk-right-about-ai-researchers-don-t-think-so). *Bloomberg News*. Retrieved 7 February 2016.
78. Elkus, Adam (31 October 2014). "Don't Fear Artificial Intelligence" (http://www.slate.com/articles/technology/future_tense/2014/10/elon_musk_artificial_intelligence_why_you_shouldn_t_be_afraid_of_ai.html). *Slate*. Retrieved 15 May 2016.
79. Radu, Sintia (19 January 2016). "Artificial Intelligence Alarmists Win ITIF's Annual Luddite Award" (https://itif.org/publications/2016/01/19/artificial-intelligence-alarmists-win-itif%E2%80%99s-annual-luddite-award). *ITIF Website*.
80. Bolton, Doug (19 January 2016). " 'Artificial intelligence alarmists' like Elon Musk and Stephen Hawking win 'Luddite of the Year' award" (https://www.independent.co.uk/life-style/gadgets-and-tech/news/elon-musk-stephen-hawking-luddite-award-of-the-year-itif-a6821921.html). *The Independent (UK)*. Retrieved 7 February 2016.
81. Garner, Rochelle (19 January 2016). "Elon Musk, Stephen Hawking win Luddite award as AI 'alarmists' " (https://www.cnet.com/news/elon-musk-stephen-hawking-win-annual-luddite-award/). *CNET*. Retrieved 7 February 2016.
82. "Anticipating artificial intelligence" (https://doi.org/10.1038%2F532413a). *Nature*. **532** (7600): 413. 26 April 2016. Bibcode:2016Natur.532Q.413. (https://ui.adsabs.harvard.edu/abs/2016Natur.532Q.413.). doi:10.1038/532413a (https://doi.org/10.1038%2F532413a). PMID 27121801 (https://pubmed.ncbi.nlm.nih.gov/27121801).
83. Murray Shanahan (3 November 2015). "Machines may seem intelligent, but it'll be a while before they actually are" (https://www.washingtonpost.com/news/in-theory/wp/2015/11/03/machines-may-seem-intelligent-but-itll-be-a-while-before-they-actually-are/). *The Washington Post*. Retrieved 15 May 2016.
84. "AI Principles" (https://futureoflife.org/ai-principles/). *Future of Life Institute*. Retrieved 11 December 2017.
85. "Elon Musk and Stephen Hawking warn of artificial intelligence arms race" (http://www.newsweek.com/ai-asilomar-principles-artificial-intelligence-elon-musk-550525). *Newsweek*. 31 January 2017. Retrieved 11 December 2017.

86. Bostrom, Nick (2016). "New Epilogue to the Paperback Edition". *Superintelligence: Paths, Dangers, Strategies* (Paperback ed.).
87. Martin Ford (2015). "Chapter 9: Super-intelligence and the Singularity". *Rise of the Robots: Technology and the Threat of a Jobless Future*. ISBN 9780465059997.
88. Grace, Katja; Salvatier, John; Dafoe, Allan; Zhang, Baobao; Evans, Owain (24 May 2017). "When Will AI Exceed Human Performance? Evidence from AI Experts". arXiv:1705.08807 (https://arxiv.org/abs/1705.08807) [cs.AI (https://arxiv.org/archive/cs.AI)].
89. 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.
90. Kaj Sotala; Roman Yampolskiy (19 December 2014). "Responses to catastrophic AGI risk: a survey". *Physica Scripta*. **90** (1).
91. Ord, Toby (2020). *The Precipice: Existential Risk and the Future of Humanity*. Bloomsbury Publishing. pp. Chapter 5: Future Risks, Unaligned Artificial Intelligence. ISBN 978-1526600219.
92. "But What Would the End of Humanity Mean for Me?" (https://www.theatlantic.com/health/archive/2014/05/but-what-does-the-end-of-humanity-mean-for-me/361931/). *The Atlantic*. 9 May 2014. Retrieved 12 December 2015.
93. Andersen, Kurt. "Enthusiasts and Skeptics Debate Artificial Intelligence" (https://www.vanityfair.com/news/tech/2014/11/artificial-intelligence-singularity-theory). *Vanity Fair*. Retrieved 20 April 2020.
94. "Tech Luminaries Address Singularity" (https://spectrum.ieee.org/computing/hardware/tech-luminaries-address-singularity). *IEEE Spectrum: Technology, Engineering, and Science News* (SPECIAL REPORT: THE SINGULARITY). 1 June 2008. Retrieved 8 April 2020.
95. http://intelligence.org/files/AIFoomDebate.pdf
96. "Overcoming Bias : I Still Don't Get Foom" (http://www.overcomingbias.com/2014/07/30855.html). *www.overcomingbias.com*. Retrieved 20 September 2017.
97. "Overcoming Bias : Debating Yudkowsky" (http://www.overcomingbias.com/2011/07/debating-yudkowsky.html). *www.overcomingbias.com*. Retrieved 20 September 2017.
98. "Overcoming Bias : Foom Justifies AI Risk Efforts Now" (https://www.overcomingbias.com/2017/08/foom-justifies-ai-risk-efforts-now.html). *www.overcomingbias.com*. Retrieved 20 September 2017.
99. "Overcoming Bias : The Betterness Explosion" (http://www.overcomingbias.com/2011/06/the-betterness-explosion.html). *www.overcomingbias.com*. Retrieved 20 September 2017.
00. Votruba, Ashley M.; Kwan, Virginia S.Y. (2014). "Interpreting expert disagreement: The influence of decisional cohesion on the persuasiveness of expert group recommendations". doi:10.1037/e512142015-190 (https://doi.org/10.1037%2Fe512142015-190).
01. Agar, Nicholas. "Don't Worry about Superintelligence" (https://jetpress.org/v26.1/agar.htm). *Journal of Evolution & Technology*. **26** (1): 73–82.
02. Greenwald, Ted (11 May 2015). "Does Artificial Intelligence Pose a Threat?" (https://www.wsj.com/articles/does-artificial-intelligence-pose-a-threat-1431109025). *Wall Street Journal*. Retrieved 15 May 2016.
03. Richard Posner (2006). *Catastrophe: risk and response*. Oxford: Oxford University Press. ISBN 978-0-19-530647-7.
04. Dadich, Scott. "Barack Obama Talks AI, Robo Cars, and the Future of the World" (https://www.wired.com/2016/10/president-obama-mit-joi-ito-interview/). *WIRED*. Retrieved 27 November 2017.
05. Kircher, Madison Malone. "Obama on the Risks of AI: 'You Just Gotta Have Somebody Close to the Power Cord' " (https://nymag.com/selectall/2016/10/barack-obama-talks-artificial-intelligence-in-wired.html). *Select All*. Retrieved 27 November 2017.
06. Clinton, Hillary (2017). *What Happened*. p. 241. ISBN 978-1-5011-7556-5. via [1] (http://lukemuehlhauser.com/hillary-clinton-on-ai-risk/)

07. "Over a third of people think AI poses a threat to humanity" (http://www.businessinsider.com/over-a-third-of-people-think-ai-poses-a-threat-to-humanity-2016-3?r=UK&IR=T). *Business Insider*. 11 March 2016. Retrieved 16 May 2016.
08. Brogan, Jacob (6 May 2016). "What Slate Readers Think About Killer A.I." (http://www.slate.com/blogs/future_tense/2016/05/06/futurography_readers_share_their_opinions_about_killer_artificial_intelligence.html) *Slate*. Retrieved 15 May 2016.
09. LIPPENS, RONNIE (2002). "Imachinations of Peace: Scientifictions of Peace in Iain M. Banks's The Player of Games". *Utopianstudies Utopian Studies*. **13** (1): 135–147. ISSN 1045-991X (https://www.worldcat.org/issn/1045-991X). OCLC 5542757341 (https://www.worldcat.org/oclc/5542757341).
10. Vincent, James (22 June 2016). "Google's AI researchers say these are the five key problems for robot safety" (https://www.theverge.com/circuitbreaker/2016/6/22/11999664/google-robots-ai-safety-five-problems). *The Verge*. Retrieved 5 April 2020.
11. Amodei, Dario, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. "Concrete problems in AI safety." arXiv preprint arXiv:1606.06565 (2016).
12. Ord, Toby (2020). *The Precipice: Existential Risk and the Future of Humanity*. Bloomsbury Publishing Plc. ISBN 9781526600196.
13. Johnson, Alex (2019). "Elon Musk wants to hook your brain up directly to computers — starting next year" (https://www.nbcnews.com/mach/tech/elon-musk-wants-hook-your-brain-directly-computers-starting-next-ncna1030631). *NBC News*. Retrieved 5 April 2020.
14. Torres, Phil (18 September 2018). "Only Radically Enhancing Humanity Can Save Us All" (https://slate.com/technology/2018/09/genetic-engineering-to-stop-doomsday.html). *Slate Magazine*. Retrieved 5 April 2020.
15. Barrett, Anthony M.; Baum, Seth D. (23 May 2016). "A model of pathways to artificial superintelligence catastrophe for risk and decision analysis". *Journal of Experimental & Theoretical Artificial Intelligence*. **29** (2): 397–414. arXiv:1607.07730 (https://arxiv.org/abs/1607.07730). doi:10.1080/0952813X.2016.1186228 (https://doi.org/10.1080%2F0952813X.2016.1186228). S2CID 928824 (https://api.semanticscholar.org/CorpusID:928824).
16. Piesing, Mark (17 May 2012). "AI uprising: humans will be outsourced, not obliterated" (https://www.wired.co.uk/news/archive/2012-05/17/the-dangers-of-an-ai-smarter-than-us). *Wired*. Retrieved 12 December 2015.
17. Coughlan, Sean (24 April 2013). "How are humans going to become extinct?" (https://www.bbc.com/news/business-22002530). *BBC News*. Retrieved 29 March 2014.
18. Bridge, Mark (10 June 2017). "Making robots less confident could prevent them taking over" (https://www.thetimes.co.uk/article/making-robots-less-confident-could-prevent-them-taking-over-gnsblq7lx). *The Times*. Retrieved 21 March 2018.
19. McGinnis, John (Summer 2010). "Accelerating AI" (http://scholarlycommons.law.northwestern.edu/cgi/viewcontent.cgi?article=1193&context=nulr_online). *Northwestern University Law Review*. **104** (3): 1253–1270. Retrieved 16 July 2014. "For all these reasons, verifying a global relinquishment treaty, or even one limited to AI-related weapons development, is a nonstarter... (For different reasons from ours, the Machine Intelligence Research Institute) considers (AGI) relinquishment infeasible..."
20. Kaj Sotala; Roman Yampolskiy (19 December 2014). "Responses to catastrophic AGI risk: a survey". *Physica Scripta*. **90** (1). "In general, most writers reject proposals for broad relinquishment... Relinquishment proposals suffer from many of the same problems as regulation proposals, but to a greater extent. There is no historical precedent of general, multi-use technology similar to AGI being successfully relinquished for good, nor do there seem to be any theoretical reasons for believing that relinquishment proposals would work in the future. Therefore we do not consider them to be a viable class of proposals."

21. Allenby, Brad (11 April 2016). "The Wrong Cognitive Measuring Stick" (http://www.slate.com/articles/technology/future_tense/2016/04/why_it_s_a_mistake_to_compare_a_i_with_human_intelligence.html). *Slate*. Retrieved 15 May 2016. "It is fantasy to suggest that the accelerating development and deployment of technologies that taken together are considered to be A.I. will be stopped or limited, either by regulation or even by national legislation."
22. McGinnis, John (Summer 2010). "Accelerating AI" (http://scholarlycommons.law.northwestern.edu/cgi/viewcontent.cgi?article=1193&context=nulr_online). *Northwestern University Law Review*. **104** (3): 1253–1270. Retrieved 16 July 2014.
23. "Why We Should Think About the Threat of Artificial Intelligence" (https://www.newyorker.com/tech/elements/why-we-should-think-about-the-threat-of-artificial-intelligence). *The New Yorker*. 4 October 2013. Retrieved 7 February 2016. "Of course, one could try to ban super-intelligent computers altogether. But 'the competitive advantage—economic, military, even artistic—of every advance in automation is so compelling,' Vernor Vinge, the mathematician and science-fiction author, wrote, 'that passing laws, or having customs, that forbid such things merely assures that someone else will.'"
24. Baum, Seth (22 August 2018). "Superintelligence Skepticism as a Political Tool" (https://doi.org/10.3390%2Finfo9090209). *Information*. **9** (9): 209. doi:10.3390/info9090209 (https://doi.org/10.3390%2Finfo9090209). ISSN 2078-2489 (https://www.worldcat.org/issn/2078-2489).
25. Domonoske, Camila (17 July 2017). "Elon Musk Warns Governors: Artificial Intelligence Poses 'Existential Risk' " (https://www.npr.org/sections/thetwo-way/2017/07/17/537686649/elon-musk-warns-governors-artificial-intelligence-poses-existential-risk). *NPR*. Retrieved 27 November 2017.
26. Gibbs, Samuel (17 July 2017). "Elon Musk: regulate AI to combat 'existential threat' before it's too late" (https://www.theguardian.com/technology/2017/jul/17/elon-musk-regulation-ai-combat-existential-threat-tesla-spacex-ceo). *The Guardian*. Retrieved 27 November 2017.
27. Kharpal, Arjun (7 November 2017). "A.I. is in its 'infancy' and it's too early to regulate it, Intel CEO Brian Krzanich says" (https://www.cnbc.com/2017/11/07/ai-infancy-and-too-early-to-regulate-intel-ceo-brian-krzanich-says.html). *CNBC*. Retrieved 27 November 2017.
28. Kaplan, Andreas; Haenlein, Michael (2019). "Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence". *Business Horizons*. **62**: 15–25. doi:10.1016/j.bushor.2018.08.004 (https://doi.org/10.1016%2Fj.bushor.2018.08.004).
29. Baum, Seth D.; Goertzel, Ben; Goertzel, Ted G. (January 2011). "How long until human-level AI? Results from an expert assessment". *Technological Forecasting and Social Change*. **78** (1): 185–195. doi:10.1016/j.techfore.2010.09.006 (https://doi.org/10.1016%2Fj.techfore.2010.09.006). ISSN 0040-1625 (https://www.worldcat.org/issn/0040-1625).
30. United States. Defense Innovation Board. *AI principles : recommendations on the ethical use of artificial intelligence by the Department of Defense*. OCLC 1126650738 (https://www.worldcat.org/oclc/1126650738).
31. Stefanik, Elise M. (22 May 2018). "H.R.5356 - 115th Congress (2017-2018): National Security Commission Artificial Intelligence Act of 2018" (https://www.congress.gov/bill/115th-congress/house-bill/5356). *www.congress.gov*. Retrieved 13 March 2020.
32. Baum, Seth (30 September 2018). "Countering Superintelligence Misinformation" (https://doi.org/10.3390%2Finfo9100244). *Information*. **9** (10): 244. doi:10.3390/info9100244 (https://doi.org/10.3390%2Finfo9100244). ISSN 2078-2489 (https://www.worldcat.org/issn/2078-2489).
33. Sotala, Kaj; Yampolskiy, Roman V (19 December 2014). "Responses to catastrophic AGI risk: a survey" (https://doi.org/10.1088%2F0031-8949%2F90%2F1%2F018001). *Physica Scripta*. **90** (1): 018001. Bibcode:2015PhyS...90a8001S (https://ui.adsabs.harvard.edu/abs/2015PhyS...90a8001S). doi:10.1088/0031-8949/90/1/018001 (https://doi.org/10.1088%2F0031-8949%2F90%2F1%2F018001). ISSN 0031-8949 (https://www.worldcat.org/issn/0031-8949).

34. Geist, Edward Moore (15 August 2016). "It's already too late to stop the AI arms race—We must manage it instead". *Bulletin of the Atomic Scientists*. **72** (5): 318–321. Bibcode:2016BuAtS..72e.318G (https://ui.adsabs.harvard.edu/abs/2016BuAtS..72e.318G). doi:10.1080/00963402.2016.1216672 (https://doi.org/10.1080%2F00963402.2016.1216672). ISSN 0096-3402 (https://www.worldcat.org/issn/0096-3402). S2CID 151967826 (https://api.semanticscholar.org/CorpusID:151967826).
35. Maas, Matthijs M. (6 February 2019). "How viable is international arms control for military artificial intelligence? Three lessons from nuclear weapons". *Contemporary Security Policy*. **40** (3): 285–311. doi:10.1080/13523260.2019.1576464 (https://doi.org/10.1080%2F13523260.2019.1576464). ISSN 1352-3260 (https://www.worldcat.org/issn/1352-3260). S2CID 159310223 (https://api.semanticscholar.org/CorpusID:159310223).
