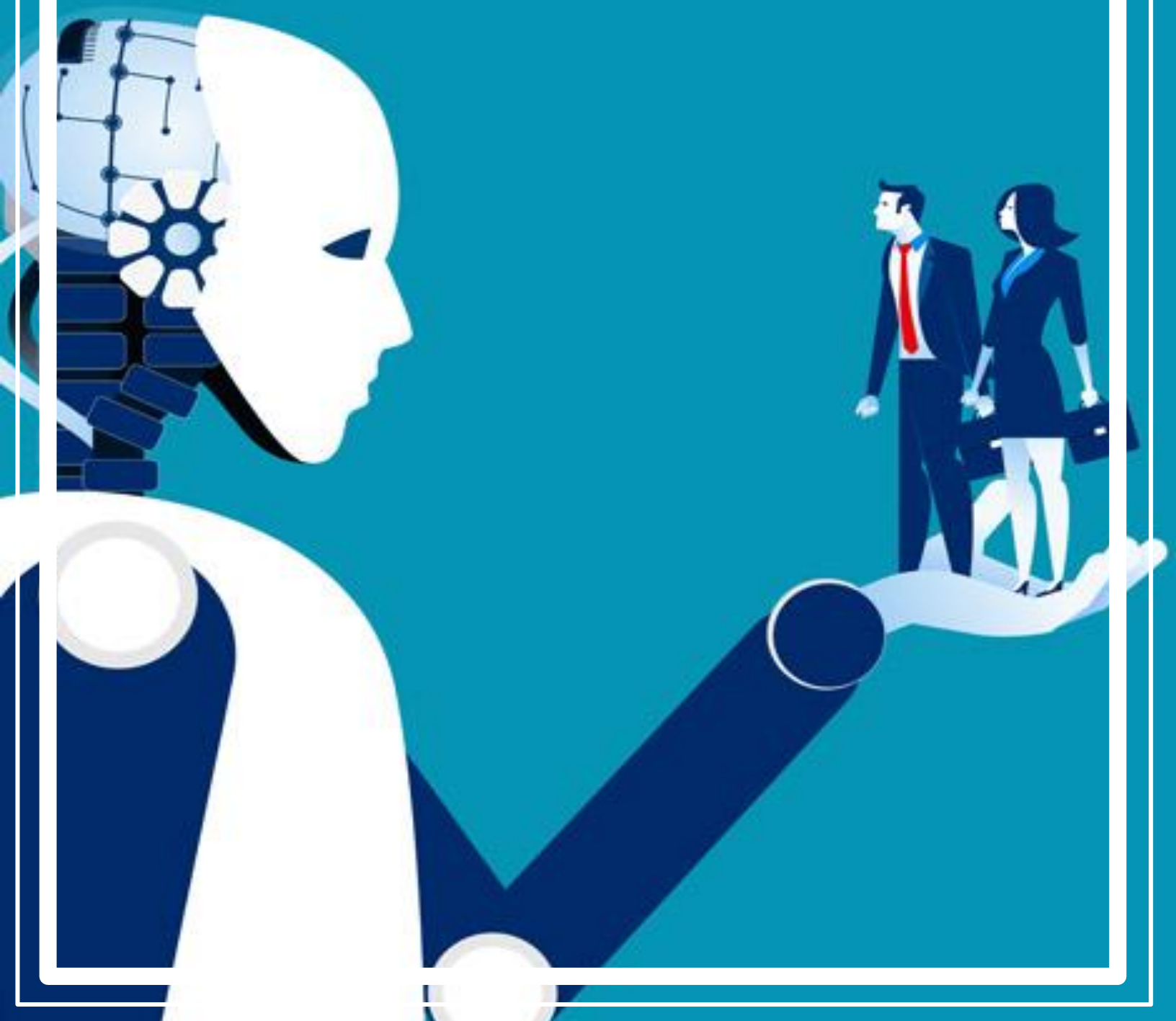
CONVENTION ON AI RIGHTS

PEAMUN XII | November 8, 2020



Dear Delegates,

Angela and I are thrilled for this committee with you guys! We hope you find this topic interesting so you can include your own ideas we can talk about together. We look forward to meeting you all, and hopefully, this will be something fun that can be a little break from everything else going on.

So to introduce ourselves, my name is Emily and I will be your chair this year. I'm a junior at PEA, and at the moment I'm in Australia, which is where I usually live. I really like cheeto puffs and I started MUN prep year, and it has become a pretty important part of my life as something I truly enjoy. My favourite part is probably getting to know and making connections with all the other delegates, so that is something we'll be trying really hard to facilitate over zoom. I hope you guys enjoy!

My name is Angela and I will be your vice chair! I'm a sophomore at PEA and I reside in New Jersey. My favorite snack is Xxtra Flamin Hot Cheetos, which Emily apparently does not prefer (though we all know which one is better). Like Emily, I started MUN my freshman year and have since enjoyed every moment of it. It has allowed me to question conventional wisdom by exploring unique topics, like the one we will be presenting for this committee. Though I wish we could meet each other in-person, I am confident our virtual conference will be fruitful!

We chose the topic of AI and AI Rights as our focus because it is clear technology is taking over the world, or at least becoming an indispensable part of our lives. Machine learning and computer programming advancements will only further the development of AI. Undoubtedly, the world will have to consider how far we can go, the ethics behind each decision, and a universal set of guidelines for AI developers to follow.

We will be focusing on three main areas: robot rights, encoding values in AI, and the limits of AI capabilities.

This background guide should serve as a thorough source for you to obtain information, and you may wish to complete additional research, for which we included suggestions for further reading.

Warmest regards,
Emily Wang & Angela Zhang

Table of Contents

Robots and Robot Rights

Artificial intelligence (AI) is a term that refers to systems that are able to mimic behaviors commonly associated with intelligent beings, such as learning and problem solving. Humanoid robots are an example of AI, a specification that we will use in this committee because the term "robot" could refer to any machine. Robotic vacuums are just one example of the machines that fall under this umbrella term, and it is clear that these robots do not resemble humans in both physical attributes and their capacity to learn.

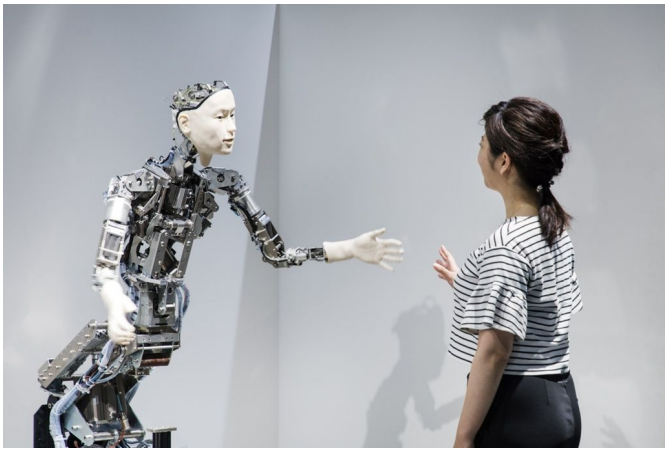
However, humanoid robots are physical replicas of human beings that are automated and contrived. One example would be care robots, which may be used to care for the elderly or disabled. Another would be companion robots, which are anthropomorphic robots that are usually used to form companionships with the elderly or single children. These are both futuristic ideas that have yet to be developed but are currently up for discussion and under the works.

These robots come with their own perks and detriments. Companionship, love, and care are all human needs. When people are unable to fulfill emotional connections in real life, perhaps they will turn to humanoid robots to cope. While this can seem like a wonderful solution to the loneliness epidemic and other such phenomena, we must consider the possibility that companionship, love, and care are unique, human needs which humanoid robots may not be able to completely satisfy.[1] If we grow too reliant on robots, our dependency may corrupt human experiences and even eliminate our relationships. However, granted that the world is able to establish a set of guidelines for AI developers to adhere to, humanoid robots may be able to

[1] https://plato.stanford.edu/entries/ethics-ai/#MainDeba

safely provide some comfort, and perhaps even utility. The following are only a few of the major issues to consider.

Due to their resemblance to humans, humanoid robots may cause individuals to grow overly attached to them. This is because humans can easily attribute humanoid characteristics to non-living objects that appear to be alive, a tendency that may establish feelings of empathy for the object. This phenomenon may make robots and AI systems appear to possess more intellectual and emotional capabilities than they truly do. It is up to this committee to examine and evaluate the dangers of this phenomenon.



There are also arguments about whether it is problematic to treat and regard humanoid in the same sense as we do humans. Discussions involve whether to use **emulation modulation** and **whole-brain emulation** methods to make humanoid robots act *even more* like humans. Opposers believe doing so will pose an **existential risk** to us, particularly because the robots are so human-like that it's even easier for there to be an **intelligence explosion.**[2] Methods of prevention for this include **boxing**, **stunting**, and **capability control methods**, all of which include some sort of limitations or restrictions in AI development to minimize the possible damage the AI can do.

[2] https://io9.gizmodo.com/can-we-build-an-artificial-superintelligence-that-wont-1501869007

If we do use emulation modulation or whole-brain emulation methods to heighten the resemblance of robots to humans (i.e., development of emotions, deviation from programmed thought), then we must also consider if robots should receive human rights. Such rights include the right to life and liberty, freedom from slavery and torture, freedom of opinion and expression, the right to work and education, and many more. This is where conflict also arises. Do AI that are made to serve humans unconditionally violate basic human rights by perpetuating slavery? Additionally, if we grant humanoid robots the freedom of opinion and expression, or the right to education, then it may interfere with value-alignment and our methods of prevention for an intelligence explosion (please see The Value-Alignment Discussion).

In addition, there is the possibility of developing robots with legal rights. The ethics behind this are doubtful but worth exploring. If AI systems or humanoid robots are given the

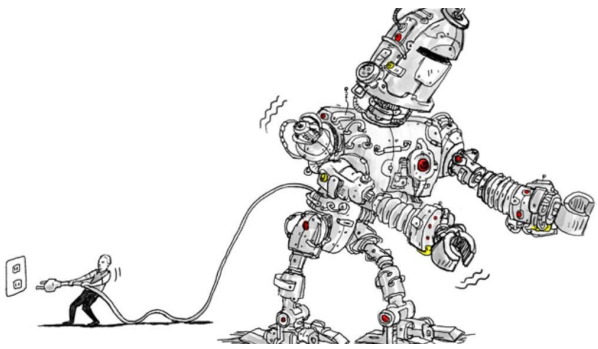
status of legal entities or anything likewise in a sense of natural persons, they can have legal rights and duties. This may allow humans to deal with robots with civil liabilities or even criminal liabilities, which is currently reserved for natural persons. Or, it is possible to only grant a certain subset of rights to humanoid robots. However, such legislative actions can have moral and legal consequences of which we must discuss.

A question to consider is if the development of humanoid robots follows **the common good principle.** If it doesn't, we must establish methods, like the aforementioned ones, to change

this, as well as research value-alignment methods. However, it is worth noting that it is not the humanoid robot itself that directly presents a threat to mankind. Rather, it is the misaligned intelligence of the humanoid robot.[3]

The Value-Alignment Discussion

As AI and computer programs are beginning to operate at increasingly greater autonomy, humans have lost the ability to keep up with and evaluate each decision that AI makes. As such, it is important to equip AI with basic principles to follow when making decisions on their own. The goal is to make sure that, when AI is operating on its own, it can automatically make "good" decisions by default, as well as understand given instructions in a context that is morally good. However, this is a difficult thing to achieve, as values may differ across different cultures and groups. It is exceedingly difficult to find a set of universal guidelines that would apply to every situation.



In 2009, the president of the Association for the Advancement of Artificial Intelligence appointed a panel of leading members to examine "the value of formulating guidelines for guiding research and of creating policies that might constrain or bias the behaviors of autonomous and semi-autonomous systems so as to address concerns." Value-alignment refers to the idea that, in order for AI to work with humans in the best way, they must have the same basic

[3] https://futureoflife.org/background/benefits-risks-of-artificial-intelligence/?cn-reloaded=1&cn-reloaded=1

values we operate with within our society. Because this is a MUN committee, we will not be focusing on the technical questions as to how to encode these values, but instead on discussing which specific principles should be universally implemented.[4]

At the moment, there seem to be three viable approaches to coming up with a set of these values. These three approaches all aim to allow us to come up with some very general principles that can be agreed on by a range of people with different beliefs and positions in society. First is the overlapping consensus, where we look for areas of widespread consensus of different groups around the world. One promising example of this type of consensus might be related to the idea of universal human rights - this is something that has been agreed on by a significant number of different countries and religions, and holds a substantial amount of support from the average human.

A second approach could be the 'veil of ignorance' idea. The approach is based upon a device proposed by philosopher John Rawls, who suggested that, when choosing principles of justice for a society, people should be asked to imagine that they do not yet know their moral views or place in this society. In doing so, these people will be more likely to choose fair values that benefit a great portion of society, in fear that they may fall into a minority group once their societal class is revealed. It is thought that this approach might work in determining principles for AI, as humans truly do not know what their position will be in an AI society or what moral views these AI hold.

[4] https://issues.org/perspective-should-artificial-intelligence-be-regulated/

The third approach is based on social choice theory, where mathematical integrations of equal individual perspectives and voting or broad discussion among individuals would ensure that ideas of which a majority of individuals can agree with and support will receive attention.

All three of these approaches seem to have the potential to develop basic principles to be implemented in AI. However, there is also the question of whether or not any principles should be enforced universally in AI. At present, AI is rapidly developing across the world, under the control of many different companies, governments, and organizations.[5] It would take substantial incentives to effectively mandate that all industry players should follow these principles. At the same time, these principles should not place too many restrictions on AI developers, as they will influence development in areas such as weapons and hospital technology. It is possible that the restricting effect of these basic principles could result in disastrous consequences in potential economic and human development.[6] During this discussion, however, it is also important to keep in mind the flip side: if we restrict the development of AI, is it ethical to let humans continue working high-risk jobs when we could develop technology to take on the task instead?[7] Should high-risk jobs, such as mine clearing, saving wounded soldiers, or saving civilians while firefighting, be allocated to robots in order to prioritise the lives of our fellow humans?

We hope that delegates will work together to draft a plan, as well as examples of basic values that should or should not be encoded in all AI. We expect discussion on both ends of whether such principles should be established.

[5] https://issues.org/perspective-should-artificial-intelligence-be-regulated/
[6] https://issues.org/perspective-should-artificial-intelligence-be-regulated/
[7] https://issues.org/perspective-should-artificial-intelligence-be-regulated/

Where To Draw The Line

As AI development has accelerated in recent years, it is estimated we will soon reach a "singularity" where AI will become so intelligent as to spiral out of our control. We will need to look back and discuss: should there be a point at which we stop this development in order to save ourselves from the effects of potentially creating machines that are so much more powerful than we could ever be?

Weapons Development

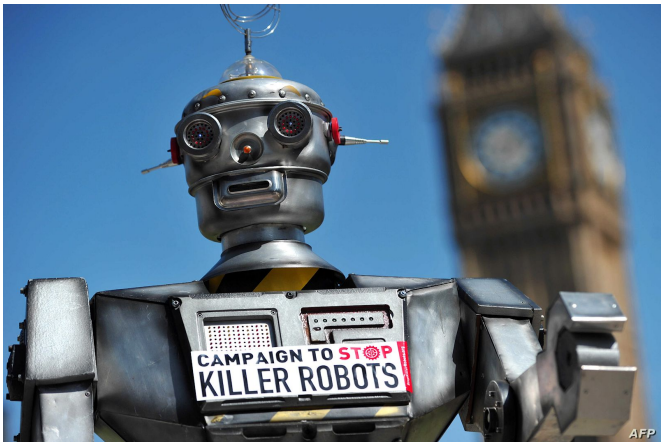
One particularly important sector we encourage you to look into is the development of autonomous weapons, which use AI to determine when to shoot and who to target in an attack. In fact, in 2015, a group of influential activists, researchers, scientists, and even businessmen submitted an open letter to the International Conference on Artificial Intelligence, calling for the UN to "ban the further development of weaponized AI that could operate beyond meaningful human control."[8] This kind of pause in the development of life-threatening technology seems sensible, and indeed such treaties in related areas such as the Treaty on the Non-Proliferation of Nuclear Weapons (1970) has had a substantial effect in reducing the amount of nuclear development worldwide. This, however, has also led to an unintended effect

[8] https://issues.org/perspective-should-artificial-intelligence-be-regulated/

of widening the gap in nuclear technology between the more powerful countries who can afford to bend the rules, and the less powerful ones who are bullied into complying. In this committee, a sub-question we will explore is if there should be a similar **treaty** directed towards **pausing the development of autonomous AI weapons**.

The CCW, or Convention on Certain Conventional Weapons, has existed since the 1980s and seeks to limit the use of certain types of weapons that are considered excessively harmful.[9] However, recent efforts to expand sections to include new types of autonomous weapons, as well as hold countries accountable to the convention, have made little progress.[10] Because of the absence of such treaties, the rapid continued development of such weapons has shown no signs of slowing. This raises the need for a new international treaty that can limit the extent of autonomous weapons development.

Factors to consider when deciding on rules of the treaty would include the potential accountability gap.[11] AI "killing machines" can create the moral issue of having robots being able to kill humans, and the consequences of an AI arms battle. The treaty should take into account methods of holding countries accountable for their efforts, as well as realizing this technology will continue to evolve and so should be broad-ranging in scope.



[9] https://www.hrw.org/news/2020/06/01/need-and-elements-new-treaty-fully-autonomous-weapons
[10] https://www.hrw.org/news/2020/06/01/need-and-elements-new-treaty-fully-autonomous-weapons
[11] https://www.hrw.org/news/2020/06/01/need-and-elements-new-treaty-fully-autonomous-weapons

AI and jobs

Another widely-shared sentiment is that AI will steal jobs from working people. Although this is true to an extent — for example, between 2000 and 2010, 1.1 million secretarial jobs and .5 million accounting and auditing jobs disappeared due to the advancement of AI abilities — it can be argued that this is more of an economic and societal issue than an AI issue. That is, in other words, the world will always be changing — it is simply a matter of providing education and support for citizens so that they have the skills to fill in new job niches generated by change.

Emotional dependency

This has already been mentioned in the previous section of this background guide, but when debating the continued development of AI without setting clear boundaries, it is also important to consider the pros of allowing humans to perceive AI as a companion, a friend, or partner. At the same time, such reliance on a robot could potentially exacerbate problems in human-to-human relationships, as the bottom line is that these robots will not, and cannot be human.

AI in Healthcare

It is easy to imagine and understand the various advantages AI can bring to the field of medicine. Today, a specialised program can diagnose skin cancer more accurately than a

certified dermatologist.[12] The telemedicine industry is thriving, with machines designed to keep patients on track at home or deliver doctors' instructions physically or over the internet. AI can operate on patients with utmost precision, much better than perhaps a tired doctor prone to mistakes.

However, there is a lack of accountability for when incorrect conclusions or analysis inevitably occur. There is the ethical issue of placing the life of a human into a robot's hands, especially within the context of "black-box" algorithms, where it cannot be logically explained how an algorithm has arrived at a given output or decision.[13] Of course, there is also the ever-present fear of when we finally integrate humans with body parts of robot systems — how will we ever draw the line?

Questions to Consider

1. Does the development of not just any robot, but humanoid robots, including companion and care robots, follow the common good principle? Is it ethical to create contrived replications of human beings?
2. Will humanoid robots fill the emotional voids among human beings, or will it only further exacerbate the problem?
3. What is the likelihood of AI posing a legitimate existential risk to humankind, and in what ways? Is an intelligence explosion probable?

---

12 https://journalofethics.ama-assn.org/article/ethical-dimensions-using-artificial-intelligence-health-care/2019-02

13 https://journalofethics.ama-assn.org/article/ethical-dimensions-using-artificial-intelligence-health-care/2019-02

4. Should there be a universal system of guidelines, restrictions, or limitations for AI developers to follow? If so, what would be entailed by this system, and who decides the principles within the system?
5. Should AI and/or humanoid robots receive human and/or legal rights?
6. Should we pause the development of autonomous AI weapons?
7. What will employment look like in the future if AI development rapidly expands?
8. Should we implement AI in healthcare, such as using AI to perform surgeries, or replacing human body parts with robot systems?

Potential Blocs

The following are some potential blocs that could form during committee. Although it is important to make sure you represent the stance of your country, there will be some leeway based on the published general plans and budget of your country. Of course, we hope you take inspiration from this and also form creative blocs on your own.

1. PRO AI AND HUMANOID ROBOTS, ANTI RIGHTS: AI is critical in advancing society. There should be no limitations or restrictions on AI's capabilities, since doing so would limit its potential and how much it can help humans. AI is unlikely to pose any legitimate threat to humans. Humanoid robots are not only beneficial, but necessary for emotional support among humans. However, humanoid robots do not need rights since they are not actually human. Delegates who could potentially hold this type of view

include China, who has plans to invest tens of billions of dollars in AI research and hopes to become the world's "leading AI innovator" by 2030.[14]

2. PRO AI, HUMANOID ROBOTS, AND RIGHTS: Same as above, but humanoid robots should be granted rights. If they are virtually replicas of humans, then it is only ethical they should be given both human and legal rights.
3. ANTI AI: AI development across all sectors must cease immediately. AI is dangerous in all aspects of life and is simply exacerbating existing problems among humans. We should seek to use our own, human capabilities rather than rely on artificial intelligence in this world. There is no need for a universal system of guidelines since AI should not be developed in the first place, as it does not follow the common good principle. Even intangible AI is a threat since all we need is a conscious internet connection to undergo recursive self-improvement.
4. ANTI HUMANOID ROBOTS: AI is of immense help to the progress of humankind and should not be restricted. We should keep developing AI like Siri, Alexa, and other intangible AI with no bodies. However, humanoid robots are unethical and are of no utility. Humanoid robots do not complete tasks humans are unable to do, which means they serve no real purpose. Plus, because they have bodies, they are capable of inflicting greater damage. We should rid them, but not other AI.
5. CONSERVATIVE AI SUPPORTER: All forms of AI should be free to be developed but only according to a universal system of guidelines. Capability control methods should be used to restrict the abilities of AI. An intelligence explosion is a legitimate threat we must

[14] https://www.unite.ai/the-different-challenges-and-approaches-to-ai-by-country/

seek to prevent. However, AI is of such benefit that we cannot just cease its development, we should only limit its abilities. An example of this type of viewpoint could be the UK, wherein a large industry is looking to implement a "comprehensive strategy for AI adoption" in order to regulate and organise this process.[15]

Further Reading

For an overview of different countries' positions on AI development:

- https://www.weforum.org/agenda/2018/09/learning-from-one-another-a-look-at-national-ai-policy-frameworks/
- https://medium.com/politics-ai/an-overview-of-national-ai-strategies-2a70ec6edfd

A recent news story on AI ethics:

- https://www.wired.com/story/google-help-others-tricky-ethics-ai/

A good site for some extra info/general questions:

- https://aiimpacts.org/

And if you feel like it, here's a report on some "AI Guidelines" that would be useful for writing resolutions in committee:

- https://link.springer.com/content/pdf/10.1007/s11023-020-09517-8.pdf

[15] https://www.unite.ai/the-different-challenges-and-approaches-to-ai-by-country/

TIMELINE

1854: George Boole argues that logical reasoning could be performed systematically in the same manner as solving a system of equations.

1898: Nikola Tesla makes a demonstration of the world's first radio-controlled vessel. The boat was equipped with, as Tesla described, "a borrowed mind."

1914: Spanish engineer Leonardo Torres y Quevedo demonstrates the first chess-playing machine, capable of king and rook against king endgames without any human intervention.

1921: Czech writer Karel Čapek introduces the word "robot" in his play R.U.R. (Rossum's Universal Robots). The word "robot" comes from the word "robota" (work).

1925: Houdina Radio Control releases a radio-controlled driverless car, travelling the streets of New York City.

1929: Makoto Nishimura designs Gakutensoku, Japanese for "learning from the laws of nature," the first robot built in Japan. It could change its facial expression and move its head and hands via an air pressure mechanism.

1950: Alan Turing proposes "the imitation game" which will later become known as the "Turing Test."

1951: Marvin Minsky and Dean Edmunds build SNARC (Stochastic Neural Analog Reinforcement Calculator), the first artificial neural network.

1952: Arthur Samuel develops the first computer checkers-playing program and the first computer program to learn on its own.

December 1955: Herbert Simon and Allen Newell develop the Logic Theorist, the first artificial intelligence program.

1956: the term "Artificial Intelligence" first coined (Dartmouth! Hanover, NH). At the time, attending scientists were extremely optimistic about the future of "solving AI", and predicted this would happen within a generation.

1957: Frank Rosenblatt develops the Perceptron, an early artificial neural network enabling pattern recognition based on a two-layer computer learning network.

1958: John McCarthy develops programming language Lisp which becomes the most popular programming language used in artificial intelligence research.

1959: Arthur Samuel coins the term "machine learning," reporting on programming a computer "so that it will learn to play a better game of checkers than can be played by the person who wrote the program.

1961: The first industrial robot, Unimate, starts working on an assembly line in a General Motors plant in New Jersey.

1961: James Slagle develops SAINT (Symbolic Automatic INTegrator), a heuristic program that solved symbolic integration problems in freshman calculus.

1964: Daniel Bobrow develops STUDENT, a natural language understanding computer program.

1965: Joseph Weizenbaum develops ELIZA, an interactive program that carries on a dialogue in English language on any topic. Weizenbaum, who wanted to demonstrate the superficiality of communication between man and machine, was surprised by the number of people who attributed human-like feelings to the computer program.

1965: Edward Feigenbaum, Bruce G. Buchanan, Joshua Lederberg, and Carl Djerassi start working on DENDRAL at Stanford University. The first expert system, it automated the decision-making process and problem-solving behavior of organic chemists, with the general aim of studying hypothesis formation and constructing models of empirical induction in science.

1966: Shakey the robot is the first general-purpose mobile robot to be able to reason about its own actions.

1968: Terry Winograd develops SHRDLU, an early natural language understanding computer program.

1970: WABOT-1, the first anthropomorphic robot, is built in Japan at Waseda University. It consisted of a limb-control system, a vision system and a conversation system.

1972: MYCIN, an early expert system for identifying bacteria causing severe infections and recommending antibiotics, is developed at Stanford University.

1973: James Lighthill reports to the British Science Research Council on the state artificial intelligence research, concluding that "in no part of the field have discoveries made so far produced the major impact that was then promised," leading to drastically reduced government support for AI research.

1974-1980: A lack of government funding and interest caused a pause in AI research. This period is often referred to as the AI winter.

1979: The Stanford Cart manages to cross a chair-filled room by itself in about five hours, becoming an early example of an autonomous vehicle.

1980: scientists at Waseda University in Japan build a musician humanoid AI “Wabot-2”, who is able to communicate with people, read musical scores, and perform tunes on an electronic organ.

1981: The Japanese Ministry of International Trade and Industry budgets $850 million for the Fifth Generation Computer project. It aimed to develop computers that could "carry on conversations, translate languages, interpret pictures, and reason like human beings."

1980s: AI work revived! British government funding (mainly to compete with Japan's Fifth Generation Computer project) kickstarted another revolution.

1986: First driverless car is a Mercedes-Benz equipped with cameras and sensors, built at Bundeswehr University in Munich. It drives up to 55mph on empty streets.

1987-1993: Another slump, mostly because of a market collapse (lack of demand for early computers), as well as following decreased funding

(although!) **1988:** Chatbot Jabberwacky is developed by Rollo Carpenter to "simulate natural human chat in an interesting, entertaining and humorous manner."

1995: chatbot A.L.I.C.E is developed by Richard Wallace, inspired by Joseph Weuzenbaum's ELIZA.

1997: IBM's Deep Blue became first computer to beat a human chess champion! It defeated Russian grandmaster Garry Kasparov.

1998: Furby, created by Dave Hampton and Caleb Chung, becomes the first pet robot.

2000: Cynthia Breazeal (MIT) develops Kismet, a robot that could recognise and simulate emotions.

2000: Honda's ASIMO, a humanoid AI, is able to walk and deliver trays to customers in a restaurant situation.

2006: The term "machine reading" is coined, defined as "an inherently unsupervised autonomous understanding of text".

2009: Northwestern University computer scientists develop Stats Monkey, an AI program that is capable of writing sports news stories without human help.

2011: IBM's Watson manages to defeat two former Jeopardy! Champions.

2014: computer "chatbot" Eugene Goostman tricked judges into thinking he was a real human during a Turing test (used to assess whether a machine is intelligent).

2014: Google's driverless car is the first to pass a U.S. state self-driving test.

March 2016: Google DeepMind's AlphaGo defeats human Go champion Lee Sedol.

Now: deep learning, artificial creativity, image understanding, natural language processing, intelligent decision-making, and physical automation are just some of the areas in which major breakthroughs are being made. However, there is no universal system of guidelines AI developers need to follow.

Glossary

(Sourced from AI Impacts, please visit aiimpacts.org for additional terminology that pique your interest)

1. **Artificial General Intelligence** (also, ***AGI***) – the intelligence of a machine that could successfully perform any intellectual task that a human being can
2. **Artificial Intelligence** (also, ***AI***) – behavior characteristic of human minds exhibited by man-made machines, and also the area of research focused on developing machines with such behavior. Sometimes used informally to refer to *human-level AI* or another strong form of AI not yet developed.
3. **Associative value accretion** – A hypothesized approach to value learning in which the AI acquires values using some machinery for synthesizing appropriate new values as it interacts with its environment, inspired by the way humans appear to acquire values

4. **Boxing** – A control method that consists of constructing the AI's environment so as to minimize interaction between the AI and the outside world.
5. **Capability control methods** – Strategies for avoiding undesirable outcomes by limiting what an AI can do
6. **Cognitive enhancement** – Improvements to an agent's mental abilities.
7. **The common good principle** – "Superintelligence should be developed only for the benefit of all of humanity and in the service of widely shared ethical ideals"
8. **Direct specification** – An approach to the control problem in which the programmers figure out what humans value, and code it into the AI
9. **Domesticity** – An approach to the control problem in which the AI is given goals that limit the range of things it wants to interfere with
10. **Emulation modulation** – Starting with brain emulations with approximately normal human motivations, and modifying their motivations using drugs or digital drug analogs
11. **Existential risk** – risk of an adverse outcome that would either annihilate Earth-originating intelligent life or permanently and drastically curtail its potential
12. **Human-level AI** – An AI that matches human capabilities in virtually every domain of interest. Note that this term is used ambiguously
13. **Human-level hardware** – Hardware that matches the information-processing ability of the human brain.
14. **Human-level software** – Software that matches the algorithmic efficiency of the human brain, for doing the tasks the human brain does.

15. **Incentive methods** – Strategies for controlling an AI that consist of setting up the AI's environment such that it is in the AI's interest to cooperate. e.g. a social environment with punishment or social repercussions often achieves this for contemporary agents
16. **Indirect normativity** – An approach to the control problem in which we specify a way to specify what we value, instead of specifying what we value directly
17. **Intelligence explosion** – A hypothesized event in which an AI rapidly improves from 'relatively modest' to superhuman level (usually imagined to be as a result of recursive self-improvement).
18. **Perverse instantiation** – A solution to a posed goal (eg, make humans smile) that is destructive in unforeseen ways (eg, paralyzing face muscles in the smiling position).
19. **Recursive self-improvement** – The envisaged process of AI (perhaps a seed AI) iteratively improving itself.
20. **Second principal-agent problem** – The emerging problem of a developer wanting their AI to fulfill their wishes.
21. **Stunting** – A control method that consists of limiting the AI's capabilities, for instance as by limiting the AI's access to information
22. **Value learning** – An approach to the value loading problem in which the AI learns the values that humans want it to pursue
23. **Value loading problem** – The problem of causing the AI to pursue human values
24. **Whole-brain emulation** – Machine intelligence created by copying the computational structure of the human brain.

Bibliography

Amitai Etzioni, Oren Etzioni. “Should Artificial Intelligence Be Regulated?” Issues in Science and Technology, February 19, 2020. https://issues.org/perspective-should-artificial-intelligence-be-regulated/.

“Arms Control Today.” Autonomous Weapons Systems and the Laws of War | Arms Control Association. Accessed August 30, 2020. https://www.armscontrol.org/act/2019-03/features/autonomous-weapons-systems-laws-war.

“Artificial Intelligence, Values and Alignment.” Deepmind. Accessed August 30, 2020. https://deepmind.com/research/publications/Artificial-Intelligence-Values-and-Alignment.

“Can We Build an Artificial Superintelligence That Won't Kill Us?” io9, January 15, 2014. https://io9.gizmodo.com/can-we-build-an-artificial-superintelligence-that-wont-1501869007.

“History of Artificial Intelligence.” Artificial Intelligence. Accessed August 30, 2020. https://www.coe.int/en/web/artificial-intelligence/history-of-ai.

“Home.” AI Impacts, February 28, 2019. https://aiimpacts.org/.

“How Far Should Artificial Intelligence Be Allowed to Go?” Future. Customer., February 12, 2019. https://www.future-customer.com/how-far-should-artificial-intelligence-be-allowed-to-go/.

Lewis, Tanya. “A Brief History of Artificial Intelligence.” LiveScience. Purch, December 4, 2014. https://www.livescience.com/49007-history-of-artificial-intelligence.html.

McFarland, Alex. “The Different Challenges and Approaches to AI by Country.” Unite.AI, April 10, 2020. https://www.unite.ai/the-different-challenges-and-approaches-to-ai-by-country/.

Moy, Glennn, Slava Shekh, Martin Oxenham, and Simon Ellis-Steinborner. “Recent Advances in Artificial Intelligence and Their Impact on Defence.” defence.gov.au. Department of Defence, 2020. https://www.dst.defence.gov.au/sites/default/files/publications/documents/DST-Group-TR-3716_0.pdf.

Müller, Vincent C. “Ethics of Artificial Intelligence and Robotics.” Stanford Encyclopedia of Philosophy. Stanford University, April 30, 2020. https://plato.stanford.edu/entries/ethics-ai/.

“The Need for and Elements of a New Treaty on Fully Autonomous Weapons.” Human Rights Watch, August 30, 2020. https://www.hrw.org/news/2020/06/01/need-and-elements-new-treaty-fully-autonomous-weapons.

Press, Gil. “A Very Short History Of Artificial Intelligence (AI).” Forbes. Forbes Magazine, December 30, 2016. https://www.forbes.com/sites/gilpress/2016/12/30/a-very-short-history-of-artificial-intelligence-ai/.

Rigby, Michael J. “Ethical Dimensions of Using Artificial Intelligence in Health Care.” Journal of Ethics | American Medical Association. American Medical Association, February 1, 2019. https://journalofethics.ama-assn.org/article/ethical-dimensions-using-artificial-intelligence-health-care/2019-02.

Rohde, Klaus, Rastko Vukovic, Michael Zeldich, Sumathy Ramesh, Jeff Hershkowitz, and Gabor Farkas. “Benefits & Risks of Artificial Intelligence.” Future of Life Institute, June

13, 2018.
https://futureoflife.org/background/benefits-risks-of-artificial-intelligence/?cn-reloaded=1.