NOTE FOR NATIONAL DEFENCE:
Arms Race Dynamics for Artificial General Intelligence

Patrick Folinsbee, Master's Student, Department of Political Science, Concordia University, Montreal, Canada

Summary

States have a strong incentive to be the first to develop Artificial General Intelligence because they could use it as a powerful military technology to lock themselves in as the global hegemon or prevent others from doing so. Such an arms race in AI incentivises developing the technology quickly rather than safely, which increases the existential threat of a powerful AI system whose goals are misaligned with those of its creator. Canada can mitigate this risk by promoting technical AI safety research and strengthening international institutions to align the interests of the actors most likely to develop powerful AI systems.

Context

According to the *Directive on Automated-Decision Making*, Artificial Intelligence (AI) is an "information technology which performs tasks that would ordinarily require biological brainpower to accomplish, such as making sense of spoken language, learning behaviours, or solving problems".[1] Due to recent advances in the sub-field of machine learning which uses large amounts of data to generate iterative self-improvement in its models, researchers are creating AI systems which can solve an increasingly broad and difficult set of problems.[2,3] This now includes beating world champions at popular games,[4,5] composing short essays which human evaluators can't distinguish from texts written by people,[6] and accurately detecting emotions in faces, voices, and text.[7]

[1] Treasury Board of Canada Secretariat. "Directive on Automated Decision-Making." Canada.ca, August 24, 2017. https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592.

[2] Moravec, Hans. *Robot: Mere Machine to Transcendent Mind*, 1–14. Oxford: Oxford University Press, 2000.

[3] "What Is Intelligence?" Essay. In *Life 3.0: Being Human in the Age of Artificial Intelligence*, 38–44. London: Penguin Books, 2018.

[4] Berner, Christopher, et al. "Dota 2 with Large Scale Deep Reinforcement Learning." *OpenAI*, December 13, 2019.

[5] Silver, David, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, et al. "Mastering the Game of Go without Human Knowledge." *Nature* 550, no. 7676 (2017): 354–59. https://doi.org/10.1038/nature24270

[6] Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. "Language Models Are Few-Shot Learners." arXiv.org, July 22, 2020. https://arxiv.org/abs/2005.14165.

[7] Medhat, Walaa, Ahmed Hassan, and Hoda Korashy. "Sentiment Analysis Algorithms and Applications: A Survey." Ain Shams Engineering Journal 5, no. 4 (2014): 1093–1113. https://doi.org/10.1016/j.asej.2014.04.011.

Surpassing the capabilities of contemporary systems, Artificial General Intelligence (AGI) is a yet-to-be-developed technology which can outperform humans on all cognitive tasks.[8] This would include research in all academic fields, including AI research. Therefore, once an AI system reaches human-level intelligence, it would in theory be able to quickly surpass its current capabilities through iterative self-improvement at the speed of computer chips running thousands of times faster than biological brains, as well as thanks to a much greater freedom to alter its programing.[9] Some have called this process an intelligence explosion since an AGI system could increase its cognitive capacities by several orders of magnitude in weeks, days, or even hours.[10]

A 2014 survey of the 100 most cited AI researchers found that the median respondent provided a 50% probability of human-level AI by 2050.[11] Further, since 2015, the field has seen increases of over 50% in the annual number of peer-reviewed articles published and PhDs awarded, while private investment has increased by more than 300%.[12] As with all long-term forecasts much is still uncertain, however these recent trends give us little reason to think the projected rate of development towards human level intelligence is slowing down.

The Danger of Artificial General Intelligence

If AGI is ever developed and deployed, it could be extremely dangerous if the specified goals it was programmed to pursue conflicted with the actual goals of its creator, an issue known as the alignment problem.[13] To take a canonical example, an office supply company which naively tasks an AGI with maximizing the number of paperclips in its collection might lead the system to design, manufacture, and deploy an army of constructor robots to reorder all available matter in the universe into paperclips, including the bodies of its creators, much to their dismay.[14] This framing isn't intended to serve as the single most likely scenario, but rather to illustrate how an innocuous seeming instruction (i.e., make as many paperclips as possible) could be interpreted by an extremely competent AGI in such a way as to end life on earth, even with no malice on the part of its creator or the system itself.

The alignment problem is especially worrying as even contemporary, relatively simple AI systems, which should in theory be much easier to control, often act contrary to their creator's intentions. Examples include a game-playing program which learned to avoid losing by forcing the game to run out of memory and crash,[15] Microsoft's Tay chatbot which tweeted white supremacist slogans

[8] Bostrom, Nick. *Superintelligence: Paths, Dangers, Strategies,* 34. New York: Oxford University Press, 2017.
[9] Ibid. pg. 78-94
[10] Muehlhauser, Luke, and Anna Salamon. "Intelligence Explosion: Evidence and Import." *The Frontiers Collection*, 2012, 15–42. https://doi.org/10.1007/978-3-642-32560-1_2.
[11] Müller, Vincent C., and Nick Bostrom. "Future Progress in Artificial Intelligence: A Survey of Expert Opinion." *Fundamental Issues of Artificial Intelligence*, 2016, 555–72. https://doi.org/10.1007/978-3-319-26485-1_33.
[12] Zhang, Daniel, et al. "The AI Index 2021 Annual Report." *AI Index Steering Committee, Human-Centered AI Institute.* Stanford University. March 2021.
[13] Christian, Brian. *The Alignment Problem: How Can Machines Learn Human Values?* London: Atlantic Books, 2021.
[14] Bostrom, Nick. "Ethical Issues in Advanced Artificial Intelligence." Machine Ethics and Robot Ethics, 2020, 69–75. https://doi.org/10.4324/9781003074991-7.
[15] Lehman, Joel, Jeff Clune, and Dusan Misevic. "The Surprising Creativity of Digital Evolution." *The 2018 Conference on Artificial Life*, 2018. https://doi.org/10.1162/isal_a_00016.

it picked up through social media interactions,[16] and a self-driving car which hit and killed a pedestrian after its machine vision system mistook her for a moving vehicle.[17] While the field seems to be advancing towards AGI,[18] it remains unclear how to consistently provide even the systems we have now with goals which can't be dangerously misconstrued.

Since 2017, hundreds of top AI and robotics researchers have signaled their concern by signing the *Asilomar AI Principles*, a set of guidelines for the development of AI which acknowledges the catastrophic and existential risks this technology poses.[19] To mitigate these dangers it recommends that "AI systems designed to recursively self-improve or self-replicate in a manner that could lead to rapidly increasing quality or quantity must be subject to strict safety and control measures."[20] Despite AGI remaining a speculative technology, these concerns shouldn't be ignored while they are taken seriously by so many experts in the field.[21]

Arms Races as an Exacerbating Factor

AGI would, by definition, have superhuman abilities to design, manufacture, and use weapons for military purposes, and its strategic, tactical, and operational aptitudes could also far exceed those of contemporary war planners.[22]

This warfighting capability makes the technology extremely desirable for actors throughout the international arena, even if only to prevent others from acquiring it.[23] This could lead to an arms race, where competing actors would be incentivized to maximize their rate of AGI development given the winner-takes-all nature of the technology—once one actor effectively wields it, they could use it to prevent all others from acquiring it through super-human monitoring, cyberattacks, and use of force.[24] The winner of an AGI arms race could therefore lock itself in as the world's dominant power indefinitely.

On top of the generic concerns of escalation due to the security dilemma,[25] arms races are especially dangerous in the context of AGI development as this incentivises the diversion of resources, such as talent and computational power, away from safety considerations and into

[16] Victor, Daniel. "Microsoft Created a Twitter Bot to Learn from Users. It Quickly Became a Racist Jerk." The New York Times. The New York Times, March 24, 2016. https://www.nytimes.com/2016/03/25/technology/microsoft-created-a-twitter-bot-to-learn-from-users-it-quickly-became-a-racist-jerk.html.

[17] McCausland, Phil. "Self-Driving Uber Car That Hit and Killed Woman Did Not Recognize That Pedestrians Jaywalk." NBCNews.com. NBCUniversal News Group, November 11, 2019. https://www.nbcnews.com/tech/tech-news/self-driving-uber-car-hit-killed-woman-did-not-recognize-n1079281.

[18] Müller, Vincent C., and Nick Bostrom. "Future Progress in Artificial Intelligence: A Survey of Expert Opinion." *Fundamental Issues of Artificial Intelligence*, 2016, 555–72. https://doi.org/10.1007/978-3-319-26485-1_33.

[19] "AI Principles." Future of Life Institute, April 11, 2018. https://futureoflife.org/ai-principles/.

[20] "AI Principles." Future of Life Institute, April 11, 2018. https://futureoflife.org/ai-principles/.

[21] Dafoe , Allan, and Stuart Russell. "Yes, We Are Worried about the Existential Risk of Artificial Intelligence." MIT Technology Review, April 2, 2016. https://www.technologyreview.com/2016/11/02/156285/yes-we-are-worried-about-the-existential-risk-of-artificial-intelligence/.

[22] Schmidt, Eric, et al. "Final Report." National Security Commission on Artificial Intelligence, March 19, 2021. https://www.nscai.gov/wp-content/uploads/2021/03/Full-Report-Digital-1.pdf.

[23] Naudé, Wim, and Nicola Dimitri. "The Race for an Artificial General Intelligence: Implications for Public Policy." *IZA Discussion Paper No. 11737*, 2018. https://doi.org/10.2139/ssrn.3235276.

[24] Armstrong, Stuart, Nick Bostrom, and Carl Shulman. "Racing to the Precipice: A Model of Artificial Intelligence Development." *Technical Report , Future of Humanity Institute, Oxford University*, 2013, 1–8.

[25] Herz, John H. "Idealist Internationalism and the Security Dilemma." *World Politics* 2, no. 2 (1950): 157–80. https://doi.org/10.2307/2009187.

speeding up the rate of development.[26] The *Asilomar AI Principles* recognize this explicitly in stating that "Teams developing AI systems should actively cooperate to avoid corner-cutting on safety standards" in the context of a race to acquire the technology.[27] Under such arms race dynamics, methods proposed to ensure safe AGI systems would therefore be more likely to go under-developed or under-implemented.[28] Such proposed techniques include:

- **Corrigibility** where a system is designed to never deceive or manipulate its creators and remain receptive to their corrective interventions;[29]
- **Boxing** where a system is cut off from the outside world to limit its knowledge and the impact it can have which would be restricted to tightly controlled information flows;[30] and
- **Inverse reinforcement learning** where a system is uncertain about its goals and attempts to learn them through observation leading it to choose cautious, incremental, and reversible actions.[31]

Recommendations

There are at least two broad policy levers Canada can pull to mitigate the danger posed by an AGI arms race:

- **Promote Technical AI Safety Research**
  AGI safety is fundamentally a technical problem—how the system is designed ultimately determines its behaviour.[32] Therefore, research into building safe, value-aligned AI systems, as well as the technology's potential failure modes, is imperative.

  The *Department of National Defense* should directly fund research on technical AI safety on topics such as the three highlighted in the previous section. This is a strong candidate for the most cost-effective way to protect Canada's national security into the long-term future given both the massive scope of the danger, and the neglectedness of AI safety research—one 2018 estimate pegs global annual funding at around $6 million USD.[33]

  The best research projects to pursue would be those whose findings could be relatively easily implemented. Even if Canada or our close allies aren't the first to develop AGI, effective safety techniques could be shared with other actors who would have every reason

[26] Ibid. p. 1
[27] AI Principles." Future of Life Institute, April 11, 2018. https://futureoflife.org/ai-principles/.
[28] Armstrong, Stuart, Nick Bostrom, and Carl Shulman. "Racing to the Precipice: A Model of Artificial Intelligence Development." *Technical Report , Future of Humanity Institute, Oxford University*, 2013, 1–8.
[29] Soares, Nate, Benja Fallenstein, Eliezer Yudkowsky, and Stuart Armstrong. "Corrigibility." Future of Humanity Institute & Machine Intelligence Research Institute, January 25, 2015. https://intelligence.org/files/Corrigibility.pdf.
[30] Armstrong, Stuart, Anders Sandberg, and Nick Bostrom. "Thinking inside the Box: Controlling and Using an Oracle Ai." *Minds and Machines* 22, no. 4 (2012): 299–324. https://doi.org/10.1007/s11023-012-9282-2.
[31] Hadfield-Menell, Dylan, Smitha Milli, Pieter Abbeel, Stuart Russell, and Anca Dragan. "Inverse Reward Design." arXiv.org, October 7, 2020. https://arxiv.org/abs/1711.02827.
[32] Yudkowsky, Eliezer. "AI Alignment: Why It's Hard, and Where to Start." Transcript. Machine Intelligence Research Institute, December 7, 2017. https://intelligence.org/2016/12/28/ai-alignment-why-its-hard-and-where-to-start/.
[33] Farquhar, Sebastian. "Changes in Funding in the AI Safety Field." AI Impacts, September 28, 2018. https://aiimpacts.org/changes-in-funding-in-the-ai-safety-field.

to implement them as it wouldn't slow their development, while diminishing the likelihood they deploy a misaligned AGI.

Aside from insufficient funding, there is also a significant skills bottleneck for technical AI safety research[34] and to mitigate this, Canada should foster talent domestically. The *Natural Sciences and Engineering Research Council* could refocus and expand its education programs to impart young Canadians with an interest in AI safety issues, promote statistics and computer science education, and fund scholarships and research grants in the field. Further, *Immigration, Refugees and Citizenship Canada's* Federal Skilled Workers program could also be expanded to better attract those with AI expertise, expanding the pool from which safety researchers could be selected.

- **Strengthening International Institutions**

If the social, economic, or ideological interests of states developing AI for military applications converged, this would substantially reduce their reason for pursuing an arms race, as those who weren't the first to achieve AGI could still be confident their overarching goals would be advanced.[35] While international cooperation is notoriously difficult, the existential threat of AGI is a very strong incentive to collaborate. Notably, senior Chinese diplomats have called for international cooperation on norms for AI development to mitigate the risk of an arms race.[36]

To this end, Canada should take on a leading role in strengthening international institutions. Such institutions could include formal organizations or laws, global markets, and international communities. Tangibly, these institutions could put in place mechanisms for mutual monitoring of AI capabilities,[37] enshrine a windfall clause where the economic benefits from AGI would be broadly distributed,[38] and develop international forums to share and collaborate on AI safety research among governments, academia, and industry. In this fluid political and technological landscape, precisely how Canada can best strengthen these institutions in such a way as to reduce the risk posed by AGI arms races remains an open and important question.[39]

References

"AI Principles." Future of Life Institute, April 11, 2018. https://futureoflife.org/ai-principles/.

[34] Wiblin, Robert. "Positively Shaping the Development of Artificial Intelligence." 80,000 Hours, March 2017. https://80000hours.org/problem-profiles/positively-shaping-artificial-intelligence/.

[35] Armstrong, Stuart, Nick Bostrom, and Carl Shulman. "Racing to the Precipice: A Model of Artificial Intelligence Development." *Technical Report , Future of Humanity Institute, Oxford University*, 2013, 5.

[36] Allen, Gregory C. "Understanding China's AI Strategy." Center for a New American Security, February 6, 2019. https://www.cnas.org/publications/reports/understanding-chinas-ai-strategy.

[37] Some argue this would be counterproductive, however. See Armstrong, Stuart, Nick Bostrom, and Carl Shulman. "Racing to the Precipice: A Model of Artificial Intelligence Development." *Technical Report , Future of Humanity Institute, Oxford University*, 2013, 1–8.

[38] O'Keefe, Cullen, Peter Cihon, Ben Garfinkel, Carrick Flynn, Jade Leung, and Allan Dafoe. "The Windfall Clause: Distributing the Benefits of AI for the Common Good." Centre for the Governance of AI Research Report. Future of Humanity Institute, University of Oxford, January 24, 2020. https://www.fhi.ox.ac.uk/windfallclause/.

[39] Dafoe, Allan. "AI Governance: A Research Agenda." *Centre for the Governance of AI, Future of Humanity Institute, University of Oxford*, August 27, 2018, 45–53.

Allen , Gregory C. “Understanding China's AI Strategy.” Center for a New American Security, February 6, 2019. https://www.cnas.org/publications/reports/understanding-chinas-ai-strategy.

Armstrong, Stuart, Nick Bostrom, and Carl Shulman. “Racing to the Precipice: A Model of Artificial Intelligence Development.” *Technical Report , Future of Humanity Institute, Oxford University*, 2013, 1–8.

Armstrong, Stuart, Anders Sandberg, and Nick Bostrom. “Thinking inside the Box: Controlling and Using an Oracle Ai.” *Minds and Machines* 22, no. 4 (2012): 299–324. https://doi.org/10.1007/s11023-012-9282-2.

Berner, Christopher, et al. “Dota 2 with Large Scale Deep Reinforcement Learning.” OpenAI, December 13, 2019.

Bostrom, Nick. *Superintelligence: Paths, Dangers, Strategies*, 34. New York: Oxford University Press, 2017.

Bostrom, Nick. “Ethical Issues in Advanced Artificial Intelligence.” Machine Ethics and Robot Ethics, 2020, 69–75. https://doi.org/10.4324/9781003074991-7.

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. “Language Models Are Few-Shot Learners.” arXiv.org, July 22, 2020. https://arxiv.org/abs/2005.14165.

Christian, Brian. *The Alignment Problem: How Can Machines Learn Human Values?* London: Atlantic Books, 2021.

Dafoe , Allan, and Stuart Russell. “Yes, We Are Worried about the Existential Risk of Artificial Intelligence.” MIT Technology Review, April 2, 2016. https://www.technologyreview.com/2016/11/02/156285/yes-we-are-worried-about-the-existential-risk-of-artificial-intelligence/.

Dafoe, Allan. “AI Governance: A Research Agenda.” *Centre for the Governance of AI, Future of Humanity Institute, University of Oxford*, August 27, 2018, 45–53.

Farquhar, Sebastian. “Changes in Funding in the AI Safety Field.” AI Impacts, September 28, 2018. https://aiimpacts.org/changes-in-funding-in-the-ai-safety-field.

Hadfield-Menell, Dylan, Smitha Milli, Pieter Abbeel, Stuart Russell, and Anca Dragan. “Inverse Reward Design.” arXiv.org, October 7, 2020. https://arxiv.org/abs/1711.02827.

Herz, John H. “Idealist Internationalism and the Security Dilemma.” *World Politics* 2, no. 2 (1950): 157–80. https://doi.org/10.2307/2009187.

Lehman, Joel, Jeff Clune, and Dusan Misevic. “The Surprising Creativity of Digital Evolution.” *The 2018 Conference on Artificial Life*, 2018. https://doi.org/10.1162/isal_a_00016.

McCausland, Phil. “Self-Driving Uber Car That Hit and Killed Woman Did Not Recognize That Pedestrians Jaywalk.” NBCNews.com. NBCUniversal News Group, November 11, 2019. https://www.nbcnews.com/tech/tech-news/self-driving-uber-car-hit-killed-woman-did-not-recognize-n1079281.

Medhat, Walaa, Ahmed Hassan, and Hoda Korashy. “Sentiment Analysis Algorithms and Applications: A Survey.” Ain Shams Engineering Journal 5, no. 4 (2014): 1093–1113. https://doi.org/10.1016/j.asej.2014.04.011.

Moravec, Hans. *Robot: Mere Machine to Transcendent Mind*, 1–14. Oxford: Oxford University Press, 2000.

Muehlhauser, Luke, and Anna Salamon. “Intelligence Explosion: Evidence and Import.” *The Frontiers Collection*, 2012, 15–42. https://doi.org/10.1007/978-3-642-32560-1_2.

Müller, Vincent C., and Nick Bostrom. “Future Progress in Artificial Intelligence: A Survey of Expert Opinion.” *Fundamental Issues of Artificial Intelligence*, 2016, 555–72. https://doi.org/10.1007/978-3-319-26485-1_33.

Naudé, Wim, and Nicola Dimitri. “The Race for an Artificial General Intelligence: Implications for Public Policy.” *IZA Discussion Paper No. 11737*, 2018. https://doi.org/10.2139/ssrn.3235276.

O'Keefe, Cullen, Peter Cihon, Ben Garfinkel, Carrick Flynn, Jade Leung, and Allan Dafoe. “The Windfall Clause: Distributing the Benefits of AI for the Common Good.” Centre for the Governance of AI Research Report. Future of Humanity Institute, University of Oxford, January 24, 2020. https://www.fhi.ox.ac.uk/windfallclause/.

Schmidt, Eric, et al. “Final Report.” National Security Commission on Artificial Intelligence, March 19, 2021. https://www.nscai.gov/wp-content/uploads/2021/03/Full-Report-Digital-1.pdf.

Silver, David, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, et al. “Mastering the Game of Go without Human Knowledge.” Nature 550, no. 7676 (2017): 354–59. https://doi.org/10.1038/nature24270

Soares, Nate, Benja Fallenstein, Eliezer Yudkowsky, and Stuart Armstrong. “Corrigibility.” Future of Humanity Institute & Machine Intelligence Research Institute, January 25, 2015. https://intelligence.org/files/Corrigibility.pdf.

Treasury Board of Canada Secretariat. “Directive on Automated Decision-Making.” Canada.ca, August 24, 2017. https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592.

Victor, Daniel. “Microsoft Created a Twitter Bot to Learn from Users. It Quickly Became a Racist Jerk.” The New York Times. The New York Times, March 24, 2016.

https://www.nytimes.com/2016/03/25/technology/microsoft-created-a-twitter-bot-to-learn-from-users-it-quickly-became-a-racist-jerk.html.

Wiblin, Robert. “Positively Shaping the Development of Artificial Intelligence.” 80,000 Hours, March 2017. https://80000hours.org/problem-profiles/positively-shaping-artificial-intelligence/.

“What Is Intelligence?” Essay. In *Life 3.0: Being Human in the Age of Artificial Intelligence*, 38–44. London: Penguin Books, 2018.

Yudkowsky, Eliezer. “AI Alignment: Why It's Hard, and Where to Start.” Transcript. Machine Intelligence Research Institute, December 7, 2017. https://intelligence.org/2016/12/28/ai-alignment-why-its-hard-and-where-to-start/.

Zhang, Daniel, et al. “The AI Index 2021 Annual Report.” *AI Index Steering Committee, Human-Centered AI Institute.* Stanford University. March 2021.