National Defence Défense nationale

Canadian Forces College

Collège des Forces Canadiennes

ETHICAL KILLER ROBOTS AND OTHER PROBLEMS

Commander Johnathan Plows

JCSP 46

Solo Flight

Disclaimer

Opinions expressed remain those of the author and do not represent Department of National Defence or Canadian Forces policy. This paper may not be used without written permission.

PCEMI 46

Solo Flight

Avertissement

Les opinons exprimées n'engagent que leurs auteurs et ne reflètent aucunement des politiques du Ministère de la Défense nationale ou des Forces canadiennes. Ce papier ne peut être reproduit sans autorisation écrite.

Canada

CANADIAN FORCES COLLEGE – COLLÈGE DES FORCES CANADIENNES

JCSP 46 – PCEMI 46

2019 - 2020

SOLO FLIGHT

ETHICAL KILLER ROBOTS AND OTHER PROBLEMS

By Commander Johnathan Plows

Word Count: 5,451

Nombre de mots : 5.451

ETHICAL KILLER ROBOTS AND OTHER PROBLEMS

Artificial Intelligence (AI) research has accelerated in recent years due to notable advancements in deep learning techniques. AI is increasingly being applied to a variety of new and sometimes novel use cases, raising a number of ethical questions for policymakers and regulators to consider. The ethical problem-space for AI suffers from conflation and confusion, compounding the challenge of creating policy to address technologies being deployed today.[1] A clear framework for understanding the problem-space of the policy debate would therefore be beneficial. The ethical problem-space of AI consists of a mix of technological factors, external societal factors, and an array of particular use cases with their own unique ethical nuances. This framework can be represented by way of a diagram (provided in Appendix) which highlights the key nodes of the problem-space. This paper will describe the key nodes in the problem-space to provide clarity in the context of an ethical debate on AI.

At the centre of the problem-space, the definition of AI has considerations unto itself, as well as one fundamental concern. This will be examined first. Then, technological factors will be discussed; in particular, recent advances in machine learning. This will lead into an examination of various societal factors in liberal western democracies. Finally, specific families of use-cases will be addressed, highlighting the unique ethical particularities of each one. This will include a description of two spectra upon which all AI uses reside. The paper is divided into subsections which each address a

[1] Johnathan Plows, "Conflation in the Artificial Intelligence Ethics Debate" (Course Paper, Toronto, Canada, Canadian Forces College, 2020).

particular node (or group of nodes) in the problem-space by providing evidence of their weight in the ethical policy debate, and highlighting their individual particularities.

The resultant framework is intended to provide clarity on the ethical issues at play in developing AI-powered technology, as well as where and how they apply. This in turn should permit a much more productive debate, and the creation of effective and relevant policy.

ARTIFICIAL INTELLIGENCE (AI)

The term AI has no consistent definition in science, policy, or law. The most commonly used definitions in academic fields incorporates the concept of the computer performing a function that hitherto required *human intelligence*.[2] This broad and potentially circular definition poses a challenge in the context of an ethical debate. Effective policy requires precision, particularly in describing applicability. In this sense, the term AI itself is a node in the problem-space.

Additionally, there is one overarching issue which pervades all potential applications of AI, which is that however effective or accurate the AI – however well it performs for a given task – humans take particular exception to a computer making a mistake that a human never would have made. It is a fundamental consideration for all AI applications that there is less tolerance for machine mistakes than human ones, particularly when the mistake is an obvious one.

[2] Edward Moore Geist, "It's Already Too Late to Stop the AI Arms Race—We Must Manage It Instead," *Bulletin of the Atomic Scientists* 72, no. 5 (September 2, 2016): 318, https://doi.org/10.1080/00963402.2016.1216672.

TECHNOLOGICAL FACTORS

Machine Learning

AI as it exists today is enabled by a range of technologies, many of them having existed for decades and being rather conventional. An AI programmed with a classical software "decision-tree" architecture poses no unique ethical considerations from a technological perspective (though, as will be described later, may still entail societal concerns, or use-case specific considerations, related to other nodes). The principle risk with conventional software is bugs, human mistakes in the code. These are well understood from an engineering ethics standpoint. However, AI research has surged in recent years due to advances in some less conventional computer technology known as deep learning. These techniques combine some long-standing *machine learning* methods with new advanced hardware, and the gold mine that is big data.[3]

Machine learning itself is a node in the problem-space. In machine learning, algorithms are used that progressively modify the behaviour of the system without human involvement.[4] This can be done in a supervised manner, where the training data is curated and labelled by humans, or in an unsupervised fashion, wherein the system may continue to learn after deployment.[5] These continuously-learning systems are subject the conflated term "emergent behaviour",[6] wherein a system might learn an unintended behaviour after

---

[3] Sherry Wasilow and Joelle B. Thorpe, "Artificial Intelligence, Robotics, Ethics, and the Military: A Canadian Perspective," *AI Magazine; La Canada* 40, no. 1 (Spring 2019): 37.

[4] J. S. Hurley, "Enabling Successful Artificial Intelligence Implementation in the Department of Defense," *Journal of Information Warfare; Yorktown* 17, no. 2 (2018): 66.

[5] National Science & Technology Council, "The National Artificial Intelligence Research and Development Strategic Plan: 2019 Update," June 2019, 24.

[6] Plows, "Conflation in the Artificial Intelligence Ethics Debate," 11. In computer science, "Emergent Behaviour" has the sense of a system learning an unintended behaviour after deployment. In biology, it refers to a group characteristic such as the intelligent flocking of birds. This leads to conflation when drone swarms are concerned (discussed later in the present paper).

deployment. In this respect, the system could develop a flaw that is not a bug, but rather, a consequence of an inherent property of the system design (being its learning characteristic). From an ethical standpoint, this is problematic in a range of use cases where some behavioural certainty is required. Many such examples are explored further down, such as weapon systems, or medical diagnostic and treatment tools.

Neural Networks

There are many different types of machine learning, but recent advances in AI development are largely attributed to a computing system called a Convolutional Neural Network (CNN).[7] When used in conjunction with a technique called backpropagation, CNN have proven to be immensely powerful in image and speech recognition, and have been critical to the advancement of the AI in self-driving vehicles, and digital assistants such as Alexa and Siri.[8] (This technique, pioneered in 2012, is now referred to as deep learning). Neural networks have certain characteristics relevant to the ethical problems-space since many AIs rely on them.

A CNN works by iteratively inputting training data to the system and then recursively feeding its own outputs back through the network until the results converge consistently on the desired output. The machinations of the CNN itself involve complex matrix mathematics, wherein weights are continually adjusted and the "logical flow" of the resulting outputs are opaque and not comprehendible by humans – the results cannot be examined to determine "why" a particular output was given.[9] The completed system is

[7] Wasilow and Thorpe, "Artificial Intelligence, Robotics, Ethics, and the Military," 37.
[8] Md Zahangir Alom et al., "The History Began from AlexNet: A Comprehensive Survey on Deep Learning Approaches," *ArXiv:1803.01164 [Cs]*, September 12, 2018, 1, http://arxiv.org/abs/1803.01164.
[9] David Gunning and David W. Aha, "DARPA's Explainable Artificial Intelligence Program," *AI Magazine; La Canada* 40, no. 2 (Summer 2019): 45.

therefore often referred to as a "black box".[10] All Neural Networks are subject to this characteristic which poses an interesting ethical challenge. In many use cases, it is difficult to trust an AI when its outputs cannot be rationally explained. Moreover, it is detrimental to the principle transparency if key decisions are made based on an unexplainable AI.

CNNs are also susceptible to a certain intriguing property of discontinuity in their output – they can be easily "fooled" by subtle perturbations of an input, imperceptible to humans.[11] In a classic example, the slight modification of an image of a school bus can produce the output "ostrich". This suggests that a malign actor could intentionally seek to undermine the proper functioning of an AI in ways unobvious to its human supervisors. These "adversarial attacks" have been the subject of much research since their discovery in 2014.[12] As described by Paul Scharre, "the AIs have weaknesses that we can't anticipate and we don't really understand how it happens or why."[13] From an engineering perspective, this is an undesirable quality for a system that requires rigorous or deterministic testing. From an ethical standpoint, it suggests that a careful and concerted approach would be required in deploying AIs with neural networks for high-stakes applications.

It is also important to consider that a given AI application, while potentially relying on multiple "black boxes", could also have many conventional software aspects of

---

[10] Gunning and Aha, 45.
[11] Christian Szegedy et al., "Intriguing Properties of Neural Networks," *ArXiv:1312.6199 [Cs]*, February 19, 2014, 1, http://arxiv.org/abs/1312.6199.
[12] Naveed Akhtar and Ajmal Mian, "Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey," *IEEE Access* 6 (2018): 14416, https://doi.org/10.1109/ACCESS.2018.2807385.
[13] Paul Scharre, *Army of None: Autonomous Weapons and the Future of War* (New York ; London: WW Norton, 2018), 188.

its overall architecture.[14] A clear understanding of how an AI was built, and where its black boxes live, would be critical to understanding the degree of ethical concern this issue might pose.

Big Data

The power of deep learning is contingent on a few things which came to a head in 2012: a landmark paper demonstrating the utility of CNN with backpropagation, the advent of affordable modern computing devices suitable to this task, and big data.[15] Deep learning requires copious amounts of training data to produce a useful and accurate model, and many large internet companies now have plenty of it. Society's increasing digital inter-connectedness is making it easier to collect more and more data for a range of applications. As Google's chief scientist put it: "We don't have better algorithms than anyone else; we just have more data."[16]

Deep learning's reliance on big data brings with it some important ethical considerations. For one, there are concerns over how and what data is collected and for what purposes (on this point there will be some overlap with societal factors). A curated set of benign images may not carry any overt privacy concerns – but the owners of the images may not wish for them to be used to train an AI. Likewise, users may not realize

[14] For instance, an image classifier would normally use an object detector to send a cropped image to an image classifier, which may in turn use multiple levels of trained algorithms to refine its response. A self-driving car would make use of machine-learning for parsing video camera imagery, but much of the "how to drive" programming would be conventional decision-tree software.
[15] Wasilow and Thorpe, "Artificial Intelligence, Robotics, Ethics, and the Military," 14.
[16] *Ibid.*, 43.

or appreciate that their Facebook interactions, or voice commands to digital assistants, are being used to train AI.[17]

Additionally, data sets do not exist in a vacuum – they are constructed from data that is available to AI developers. In this regard, they are concerningly subject to societal biases, and skewed towards demographics that produce more data. Recently, an AI trained to recognize race and gender was noticeably more accurate at identifying white males than other races, and least effective at identifying women with darker skin.[18] This sort of bias pervades many AI use cases, and is an unintended consequence of what data is available, not necessarily the designer's own biases. It is an important ethical consideration that AI data sets must be carefully constructed, and the results tested for unintended bias.

Finally, the reliance on the "bigness" of the data means that statistical edge-cases will invariably be missed, resulting in an AI which renders incorrect outputs when these arise. Where a human may have sufficient insight to recognize a unique set of conditions and render appropriate judgement, a machine-learned AI will always be limited by the data on which it was trained. As will be discussed later, law and medicine present such edge-cases with some regularity.[19] This limitation of AI in certain high-stakes applications is an important ethical consideration in determining the extent of their use, and to what degree they are relied upon for accuracy.

---

[17] Wasilow and Thorpe, 42.
[18] *Ibid.*, 43.
[19] Seumas Miller, "Machine Learning, Ethics and Law," *Australasian Journal of Information Systems* 23, no. 0 (May 1, 2019): 7, https://doi.org/10.3127/ajis.v23i0.1893.

Artificial General Intelligence (AGI)

Thus far, only so-called "weak" AI has been discussed. Another form, called "strong" AI, or Artificial General Intelligence (AGI), is contemplated by science to represent a machine with cognition equivalent to that of a living being – one that can reason and think – or at least appears to in any way that matters.[20] This distinct and much smaller field of research focusses on replicating either the psychological processes of the mind, or the structure of the brain.[21] Such a novel form of AI would certainly open its own Pandora's box of ethical considerations. A popular source of distress would be an eventual singularity, where machines could surpass human intelligence.[22] A dire prognosis of such an occurrence could be the extermination of the human race by machines, as in the film *Terminator*.[23]

It is important to note that current technology is a long way from producing anything close to AGI. Nevertheless, an effective ethical discussion in today's context must clearly differentiate these two forms and establish a clear scope for what is under consideration. Conflation in the debate-space suggests these differentiations are not always made, resulting in potential confusion in the ethical debate.[24]

---

[20] Pei Wang, "On Defining Artificial Intelligence," *Journal of Artificial General Intelligence; Vienna* 10, no. 2 (2019): 15, http://dx.doi.org.cfc.idm.oclc.org/10.2478/jagi-2019-0002.
[21] *Ibid.*, 4.
[22] Adriana Braga and Robert K. Logan, "AI and the Singularity: A Fallacy or a Great Opportunity?," *Information* 10, no. 2 (February 1, 2019): 1, https://doi.org/10.3390/info10020073.
[23] James Johnson, "Artificial Intelligence & Future Warfare: Implications for International Security," *Defense & Security Analysis* 35, no. 2 (April 3, 2019): 159, https://doi.org/10.1080/14751798.2019.1600800.
[24] Plows, "Conflation in the Artificial Intelligence Ethics Debate," 3.

EXTERNAL SOCIAL FACTORS

Liberal Democracy and Society

This paper largely concerns itself with the ethical considerations of the West, since it is within this framework that AI policy is being crafted in Canada and amongst its partners. In a liberal democracy, ethics are freely debated, and individual beliefs and opinions matter. Policy and law typically aim to strike a balance between Kantian rights-based value systems, and classical utilitarianism (providing the most good for the most people, or its converse, the least evil).[25] It is from these political traditions that the word *autonomy* has its roots – representing the right to self-government and liberty – and to an extent, independent moral agency.[26] This historical context is relevant when discussing Autonomous Systems (AS), which are generally AI-controlled. The autonomy of a system is meaningfully divorced from that of a society, yet in a free ethical debate, society's perceptions are important, and machine autonomy raises many concerns for constituents.

In this regard, many activist groups, think tanks, and civil society organizations, have taken up advocacy in the ethical debate concerning AI. Many organizations have offered policy recommendations; among them, Amnesty International's "Toronto

[25] Ronald C. Arkin, "Ethics and Autonomous Systems: Perils and Promises [Point of View]," *Proceedings of the IEEE* 104, no. 10 (October 2016): 1779, https://doi.org/10.1109/JPROC.2016.2601162.
[26] Amitai Etzioni and Oren Etzioni, "Incorporating Ethics into Artificial Intelligence," *The Journal of Ethics; Dordrecht* 21, no. 4 (December 2017): 410, http://dx.doi.org/10.1007/s10892-017-9252-2.

Declaration"[27] and the independently-convened "Montreal Declaration."[28] The qualms of individuals are also relevant. Prominent computer scientist Stuart Russel has co-founded the "Campaign to Stop Killer Robots", a grassroots movement opposing the use of AI in the intentional application of lethal force.[29] As a researcher in the field, he has (and is entitled to) his own scruples concerning how his work is used, as do other developers and practitioners.

In public opinion, futurism and fearmongering can often dominate the narrative, with reason being abandoned to Hollywood sensationalism.[30] Google recently had to withdraw its support to the Pentagon's *Project Maven* due to a vocal backlash and walk-out of its employees opposing the company's involvement in a military AI project.[31] These societal voices are important and relevant in the policy debate.

Societal Norms and Privacy (Norms)

Underlying western culture is the notion of "folk morality", which largely amounts to what "the average person would deem moral or acceptable". It is this folk morality which AI practitioners and policymakers normally aim to strike; though, it is

---

[27] Amnesty International, "The Toronto Declaration: Protecting the Rights to Equality and Non-Discrimination in Machine Learning Systems," *Access Now* (blog), May 16, 2018, https://www.accessnow.org/the-toronto-declaration-protecting-the-rights-to-equality-and-non-discrimination-in-machine-learning-systems/.

[28] Christophe Abrassart et al., "Montréal Declaration for a Responsible Development of Artificial Intelligence," 2018, https://5dcfa4bd-f73a-4de5-94d8-c010ee777609.filesusr.com/ugd/ebc3a3_506ea08298cd4f8196635545a16b071d.pdf.

[29] "The Campaign To Stop Killer Robots," accessed November 20, 2019, https://www.stopkillerrobots.org/about/.

[30] Arkin, "Ethics and Autonomous Systems," 1779.

[31] Johnson, "Artificial Intelligence & Future Warfare," 156.

[32] Where the supplied diagram uses an abridged or alternate node name, the associated node name from the diagram is included in brackets within the subtitles of the main body of this text for clarity

itself nebulously defined and subject to debate.[33] This folk morality is culturally influenced, and in the context of AI, heavily swayed by societal norms.

Here, privacy takes a center stage along three different thrusts. First, western society's interconnectedness through social media has brought about new conceptions of privacy altogether.[34] Second, AI-enabled systems can infer personal details using data on-hand that previously would have remained private: examples include Facebook inferring and tagging users' political alignment based on their interactions with the platform,[35] as well as identifying which of the users are at risk of suicide.[36] Third, a reverse-privacy argument exists in some applications wherein users are more comfortable with an AI accessing their data (such as a corporate email usage policy verification) than they would be with other human beings.

It is notable that many societies may currently be at a privacy inflexion point as they consider deploying contact-tracing apps in mobile devices to help control the COVID-19 pandemic.[37]

---

[33] George R. Lucas, *Engineering, Ethics, and Industry: The Moral Challenges of Lethal Autonomy* (Oxford University Press, 2013), 222, https://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780199926121.001.0001/acprof-9780199926121-chapter-10.

[34] Wasilow and Thorpe, "Artificial Intelligence, Robotics, Ethics, and the Military," 42.

[35] Jeremy B. Merrill, "Liberal, Moderate or Conservative? See How Facebook Labels You," *The New York Times*, August 23, 2016, sec. U.S., https://www.nytimes.com/2016/08/24/us/politics/facebook-ads-politics.html.

[36] "Facebook Increasingly Reliant on A.I. To Predict Suicide Risk," *All Things Considered* (NPR, November 17, 2018), https://www.npr.org/2018/11/17/668408122/facebook-increasingly-reliant-on-a-i-to-predict-suicide-risk.

[37] Luca Ferretti et al., "Quantifying SARS-CoV-2 Transmission Suggests Epidemic Control with Digital Contact Tracing," *Science*, March 31, 2020, 4, https://doi.org/10.1126/science.abb6936.

Policy and Law

All these societal factors play into burgeoning AI policy and legislation. Policies currently exist at various corporate and governmental levels: notable examples include the Canadian Treasury Board "Directive on Automated Decision Making",[38] and the United States Department of Defence (US DoD) "Directive on Autonomy in Weapon Systems".[39] There remains very little in the way of legislation concerning AI directly, but the recent "General Data Protection Regulation" (GDPR) in Europe, and other Western privacy laws, are beginning to be adapted to current technological reality.[40] While policy and law are necessarily derived in part from the ethical debate, they in turn also influence the development, interpretation, and application of further policies and laws.

Research and Development by the Private Sector

Finally, technological and societal factors share one node in common in the ethical problem-space, and that is that innovations in AI are almost entirely being borne by the private sector, outside of government or defence contracts. In the defence context, this is referred to as dual-use technology, and it poses its own ethical considerations. AI is rapidly being developed and adopted in a range of non-military applications.[41] With private industry leading the effort rather than being driven by the requirements of a

---

[38] Treasury Board Secretariat, "Directive on Automated Decision-Making" (Government of Canada, February 5, 2019), https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592#appA.

[39] Department of Defense, "Autonomy in Weapon Systems," Directive, 3000.09, May 8, 2017, https://www.esd.whs.mil/Portals/54/Documents/DD/issuances/dodd/300009p.pdf.

[40] Saswat Sarangi and Pankaj Sharma, *Artificial Intelligence: Evolution, Ethics and Public Policy* (Milton, UNITED KINGDOM: Taylor & Francis Group, 2018), 102, http://ebookcentral.proquest.com/lib/cfvlibrary-ebooks/detail.action?docID=5517469.

[41] Johnson, "Artificial Intelligence & Future Warfare," 154.

responsible government, policymakers are lagging the deployment of AI today, potentially to the detriment of governmental goals.

AI IN PRACTICE – CONSIDERATIONS PARTICULAR TO USE CASES

The Use-Case Spectra

The following sections concern the unique considerations of various categories of AI use-cases. While these applications are wide-ranging, each potential usage can be conceived as existing somewhere along two spectrums. The first spectrum ranges from mundane to high-stakes usages. Easily conceptualized, an AI which curates a music playlist for a willing listener, or governing the watering of domestic plants, pose few novel ethical concerns. By contrast, an AI with a measure of influence over whether a person lives or dies has significant call for the application of moral agency in its use. Since this paper concerns the ethical problem-space for AI, these sections will primarily address applications nearer the high-stakes end of this spectrum.

Similarly, all AI applications exist on a spectrum from low-fidelity "decision aids" to fully Autonomous Systems (AS) able to directly influence the physical world in some way. For instance, an AI could be employed in assisting a doctor with a medical diagnosis, but be relegated as a mere factor among several others for the doctor to consider. This application remains high-stakes based on its potential impact to human life, but its influence is largely restrained – the AI is a tool supplementing human cognitive function. In contrast, the AI in a self-driving vehicle wields significant control over a large moving object and could directly threaten human lives. It must be noted that there is a significant amount of variation in the degree of autonomy a system possesses, with no

clear threshold between a highly-reliant decision aid, and a weakly autonomous system.[42] The nodes discussed in the following sections refer to general use cases, each of which – in their own right – have a range of potential applications anywhere on this latter spectrum.[43]

Criminality and Illegal Use

There are many concerns over the power of AI being used for illegal activities and organized crime. Recent examples include the Volkswagen emissions cheating scandal, and Uber's use of a tool called Greyball (in cities where Uber is illegal) to detect when a ride-hailer is likely to be a police officer.[44] Both these schemes involved the use of AI. It is also easily envisaged that criminals could use AI to target victims vulnerable to fraud.

As mentioned earlier, many AI systems have unintuitive frailties that are still being discovered. There are significant concerns that these could be exploited by malicious actors to defeat the proper functioning of a system or safety feature; such as by veering a self-driving vehicle off-course, or causing the arrest of an innocent person. The new and novel ways AI weaknesses could be exploited pose ethical considerations that challenge conventional conceptualizations of criminal behaviour.

AI could also be used in police work, such as assisting in profiling criminals for crimes under investigation, or even, to prevent them.[45] They could also be used to

---

[42] Irving Lachow, "The Upside and Downside of Swarming Drones," *Bulletin of the Atomic Scientists* 73, no. 2 (March 4, 2017): 97, https://doi.org/10.1080/00963402.2017.1290879.

[43] The reader should infer that the relative positions of these nodes on the diagram do not correlate to a particular location on the spectrum. Each individual node, regardless of colour or position, could have implementations anywhere on the spectrum from a decision aid to an AS.

[44] Danton S. Char, Nigam H. Shah, and David Magnus, "Implementing Machine Learning in Health Care -- Addressing Ethical Challenges," *The New England Journal of Medicine; Boston* 378, no. 11 (March 15, 2018): 982, http://dx.doi.org.cfc.idm.oclc.org/10.1056/NEJMp1714229.

[45] Miller, "Machine Learning, Ethics and Law," 4.

monitor the activities of individuals in privileged occupations, such as government officials, military personnel or law enforcement.[46] In such circumstances, what thresholds would be required to trigger an investigation or punitive actions?

Corporate Use

Another recurring theme in the AI policy debate is the notion of trusting corporations to be ethical custodians of the technology and its associated data. The quintessential example would be that of Cambridge Analytica, which utilized an AI with Facebook data to non-trivially interfere in the 2016 US election. There is overwhelming consensus that this was an undesirable, even reprehensible, activity; but it has not yet been determined whether any actual laws or policies were broken.[47] The current legal case relies heavily on whether or not the company obtained its data legally, and not what it actually did with it, or how it employed its AI. It is clearly an important ethical consideration in policymaking that capitalist entities are not governed by a prominent social conscience and therefore it is the law which must protect society from AI harms.

The Human Experience (Human+)

Society's direct relationship with AI poses some interesting ethical considerations as well. There is no shortage of research on how computing, social media, and advanced technology have impacted the human experience over the last few decades. This impact will be further heightened by robotics, digital assistants, and other powerful AI available in the palms of our hands. People's relationships with technology is changing their

---

[46] Miller, "Machine Learning, Ethics and Law," 1.

[47] Jim Isaak and Mina J Hanna, "User Data Privacy: Facebook, Cambridge Analytica, and Privacy Protection - IEEE Journals & Magazine," *IEEE Computer* 51, no. 8 (August 2018): 57, https://doi.org/10.1109/MC.2018.3191268.

relationships with each other. Social interactions are more and more occurring over digital means, and children are growing up in homes where it is common to interact with Siri or Alexa – forming relationships with digital assistants.[48] These interactions are singularly different from how people interact with other humans and could impact how they treat each other. There is also growing concern that intimate robotics, already available to consumers, could be damaging for individuals or society at large. This is a market which policy and law have yet to address.[49] The adult industry is often said to be the pioneer driving the adoption of now-mainstream technologies, such as VHS and digital video.[50] When systems are designed specifically to nurture attachment and emotional responses from humans, this poses novel ethical considerations for how AI will shape the human experience.

Livelihood and Welfare

Institutional use cases also pose important considerations, particularly those that involve decisions which will impact the welfare, freedom, or livelihood of people. Two prominent examples will be discussed, medicine and law, because these are fields where AIs are already being employed as decision aids.

Medical research brings in big grants and has long been at the forefront of technological innovation. AI-enabled diagnostic tools are soon expected to surpass human doctors in fields such as anatomical pathology and radiology.[51] In law, some US judges

---

[48] Steven Brykman, "Douglas Adams Was Right: 'Genuine People Personalities' Are Coming to Our Gadgets," Ars Technica, December 22, 2018, https://arstechnica.com/gadgets/2018/12/douglas-adams-was-right-get-ready-to-talk-with-digital-personalities/.
[49] Arkin, "Ethics and Autonomous Systems," 1780.
[50] *Ibid.*, 1781.
[51] Char, Shah, and Magnus, "Implementing Machine Learning in Health Care -- Addressing Ethical Challenges," 981.

are using AI to assist in meting sentences. In some jurisdictions, civil courts have used them to determine divorce settlements in circumstances where both parties agree with the determination of the AI – sparing them lengthy and costly legal proceedings.[52]

These applications are critically subject to the technological limitations discussed earlier. AI-derived judgements in some jurisdictions have been shown to exhibit racial bias, because the training data (sentences imposed by human judges) also contained these biases.[53] Similarly, medical AI tools could fall short where certain genetic populations are under-represented in research data. Privacy and big data considerations are also prominent in medicine, where doctor-patient confidentiality is a tenet of the Western practice. A machine-learned AI would require doctors to disclose medical records, contravening this principle as ingrained as their Hippocratic oath.[54] Additionally, both fields suffer the risk of lacking subtlety in statistical edge-cases: in medicine, this could result in the withdrawal of care where a doctor might reasonably have recognized a chance of survival;[55] or similarly in law, a wrongful conviction.[56]

These use cases have exceptional considerations as well. For instance, in both medicine and law, the bulk of the training data available consists of past decisions and diagnoses made by highly skilled humans. Once these functions are relegated to the AI, human skills could atrophy, and the source of "new" training data dries up. In some circumstances, the AI would then never improve, and would permanently amplify any

---

[52] Miller, "Machine Learning, Ethics and Law," 5.
[53] Char, Shah, and Magnus, "Implementing Machine Learning in Health Care -- Addressing Ethical Challenges," 982.
[54] Char, Shah, and Magnus, 983.
[55] *Ibid.*, 982.
[56] Miller, "Machine Learning, Ethics and Law," 7.

mistakes or misjudgements inherent in its original training. In medicine, concern abounds that data-driven research is replacing the "collective medical mind", wherein previously it was individual clinical experience.[57]

In AI applications where human livelihood or welfare are concerned, there has normally been a trained and experienced human making determinations. The subjugation of human judgment to algorithms and AI clearly raises intriguing ethical considerations, particularly when facing the prospect of extinguishing the very source of their human-generated training data.

Warfare

The paper will now focus on use cases with lethal potential, beginning with warfare in general, followed by subnodes of individual significant weight. AI applications in warfare are nearly as broad as in the clandestine domains, but include some particularities such as in intelligence collection and analysis.[58] Warfare is a rich domain for ethical debate and the introduction of AI brings with it many high-level concerns.

For instance, there is growing concern that the use of AI in warfare will spur an AI arms race with competitors.[59] Certain AI applications, such as in cyber offense, are difficult to attribute to state actors and could exacerbate Grey Zone competition.[60] AI on the front lines of interstate competition could lower the barriers to entry into conflict, and result in more escalation into lethal warfare.[61] Recent advances in AI are also

---

[57] Char, Shah, and Magnus, "Implementing Machine Learning in Health Care -- Addressing Ethical Challenges," 982.
[58] Hurley, "Enabling Successful Artificial Intelligence Implementation in the Department of Defense," 64.
[59] Johnson, "Artificial Intelligence & Future Warfare," 156.
[60] *Ibid.*, 153.
[61] Wasilow and Thorpe, "Artificial Intelligence, Robotics, Ethics, and the Military," 39.

democratizing the technology, making it accessible to non-state actors. The proliferation of AI could enable novel uses by insurgents for engaging in asymmetric warfare.[62]

Even when used as a decision aid, there exists concerns over liability for AI-adjacent human-made decisions with lethal consequences, whether the AI's advice was followed or not.[63] In the case of autonomy, or near-autonomous systems, the very nature of command is challenged.[64]

There are even concerns that AI could destabilize global nuclear peace by enabling sensors that would render the oceans transparent, and the pinnacle of nuclear deterrence – the nuclear submarine – would lose its all-important stealth ability.[65] Finally, there have been many calls to ban certain AI applications in warfare, such as has been done with land mines and chemical weapons. These efforts suffer the same definitional issues previously mentioned, and bans remain difficult to enforce.[66]

AI in the Kill-Chain

A key subnode of warfare concerns the use of AI in the sequence of events that results in lethal or destructive force; sometimes referred to as the kill-chain. Examples could include robot soldiers (not yet known to exist in any fully-autonomous form), or AI-controlled aircraft that patrol sectors and prosecute targets with specific characteristics, such as the Israeli Harpy – a so-called loitering munition.[67] It could

---

[62] Johnson, "Artificial Intelligence & Future Warfare," 153.
[63] Wasilow and Thorpe, "Artificial Intelligence, Robotics, Ethics, and the Military," 40.
[64] Ross Pigeau and Carol McCann, "Re-Conceptualizing Command and Control," *Canadian Military Journal* 3, no. 1 (2002): 61.
[65] Geist, "It's Already Too Late to Stop the AI Arms Race—We Must Manage It Instead," 319.
[66] *Ibid.*, 320.
[67] Johnson, "Artificial Intelligence & Future Warfare," 151.

equally include sensor-enabled AI decision-aides of various fidelity, such as the naval AEGIS system, or the Patriot air defence system, wherein humans exert varying levels of control over the final decision to fire.[68]

Such implementations are often termed "human in the loop" or "human on the loop", meaning that a human decision is required for action in the former case, or that a human is supervising with veto powers in the latter. In the case of weapon systems, there is near consensus in the United Nations (UN) that these should remain under "meaningful human control"[69], or in Canada's case, the arguably more permissive "appropriate human involvement".[70] Both forms are subject to interpretation and debate.

Ethical arguments in this node principally rely on the Law of Armed Conflict (LOAC) to determine whether a particular weaponized AI meets the requirements of proportionality, discrimination, military necessity, and the avoidance of needless suffering. Policymakers must also consider the "battle for the hearts and minds" of their adversary and the international community – it would not serve a Western nation's objective to use weapons considered reprehensible by their allies or competitors, even if they had determined them to be ethical.[71]

Lethal Autonomous Weapon Systems (LAWS)

Lethal Autonomous Weapon Systems (LAWS) are the edge case in the kill chain debate, where a weapon could be considered to operate in the absence of "meaningful" human control. Here, the spectrum of autonomy becomes particularly poignant. Loitering

---

[68] Lucas, *Engineering, Ethics, and Industry*, 220.
[69] Arkin, "Ethics and Autonomous Systems," 1780.
[70] Wasilow and Thorpe, "Artificial Intelligence, Robotics, Ethics, and the Military," 46.
[71] *Ibid.*, 41.

munitions or robot soldiers are the most obvious examples of LAWS. There are, however, many unobvious examples in the ethical debate. Some would consider the acoustic homing torpedo (c. Second World War) as the first autonomous weapon.[72] What constitutes LAWS from a regulatory or policy perspective remains to be determined. The issue of weapons autonomy is best characterized by Paul Scharre as the system's "freedom in time and space" to manoeuvre and select targets.[73] A homing torpedo is fired by a human into a relatively narrow maneuvering bracket and lifespan, and recognizes ships by acoustic or magnetic signatures. In contrast, the Harpy II has a wide operating area and six-hour flight time. It discriminates targets based on electromagnetic signatures stored in its library.[74]

LAWS are a current hot button in the warfare and kill-chain ethical debates. Notwithstanding their utility for military objectives, there are arguments on either side, including considerations as to whether a robot soldier could in fact be more ethical on the battlefield and better able to enforce the rules of engagement or LOAC.[75] Stuart Russel, on the other hand, envisions an "undesirable" future where slaughterbots roam in swarms, programmed to deliver lethal explosives into human skulls.[76]

Swarms and Drone Warfare

Drone swarms are a key subnode under warfare. Drone swarms are currently enabled by machine learning among other AI-related technologies. In the concerns over an AI arms race, swarms figure prominently since modeling has demonstrated the best

---

[72] Geist, "It's Already Too Late to Stop the AI Arms Race—We Must Manage It Instead," 2016.
[73] Scharre, *Army of None*, 52.
[74] Johnson, "Artificial Intelligence & Future Warfare," 151.
[75] Lucas, *Engineering, Ethics, and Industry*, 225.
[76] Stuart Russell, "Ethics of Artificial Intelligence," *Nature; London* 521, no. 7553 (May 28, 2015): 416.

defence against a drone swarm may in fact be another drone swarm.[77] There are repeat considerations here, such as privacy (when used for surveillance) and proportionality in lethal warfare. However, there are some unique complications in the ethical debate, owing in part to the conflation of terms in this domain of research. Foremost of these is "emergent behaviour", which in biology is most used to describe the intelligent flocking of birds (not unlike a swarm). As mentioned earlier, it has a different meaning in computer science. When applied to swarming technology, it conjures the notion that a swarm could develop unprogrammed behaviour.[78] While this would be true if a swarm were deployed in an unsupervised learning state – it would hardly be the case for all swarming applications – a point that appears to be missed in the current state of the policy debate.[79]

Confusion in the policy debate concerning drone swarms also draws in arguments largely unrelated to AI because of their nexus with drone warfare writ large. A remotely piloted drone may not involve AI, and yet the AI ethics debates concerns itself with this equally contentious practice with near equal fervour.[80] Remotely piloted vehicles have been controversial in warfare since their inception. Concerns include privacy in persistent surveillance, their use in "black ops" targeted killing, the status of their human pilots as combatants, and numerous others.

---

[77] Lachow, "The Upside and Downside of Swarming Drones," 100.
[78] Wasilow and Thorpe, "Artificial Intelligence, Robotics, Ethics, and the Military," 40.
[79] Plows, "Conflation in the Artificial Intelligence Ethics Debate," 11.
[80] Lucas, *Engineering, Ethics, and Industry*, 212.

Safety-Critical and Self-Driving Vehicles

The final major node concerns safety-critical systems, where an AI governs a system with the potential to cause harm, but is not intended to. This debate is dominated by self-driving vehicles.[81] Autonomous cars have big profit potential for corporations and are therefore a significant research area. On the one hand, "humans are the most dangerous things on the road"[82] and traffic accidents are the leading cause of death in many parts of the world. On the other, AI drivers have not yet been shown to be safer than humans, and have been involved in high-profile fatalities with both Uber and Tesla vehicles. Liability for AI-involved deaths is currently being decided in courtrooms in the absence of specific legislation concerning such incidents,[83] and as the old adage goes, "hard cases make bad law".[84]

The ethical debate for cars concerns itself also on how the vehicle should be programmed to drive. Should it emulate imperfect human driving, or follow the letter of the law precisely? Or, should new driving laws be established specific for self-driving cars? It also concerns itself with the classic "trolley problems" in which the AI must decide between two distinct negative outcomes: such as either harming the passengers, or harming pedestrians; because conceivably, the AI could determine the inevitability of either.[85] These scenarios tend to be highly contrived.[86] Real-world situations would seldom present so definitive and binary a range of outcomes. Nevertheless, these trolley

---

[81] Arkin, "Ethics and Autonomous Systems," 1780.
[82] *Ibid.*
[83] Arkin, 1780.
[84] Etzioni and Etzioni, "Incorporating Ethics into Artificial Intelligence," 416.
[85] *Ibid.*, 415.
[86] *Ibid.*

problems make up a significant portion of the ethical debate because there is no firm agreement in the folk morality on the right course in any given scenario.[87] In a regular traffic incident, there could be a broad range of actions a human might take that would be considered morally or legally acceptable. However, AI tends to be held to a higher moral standard than are humans.[88] Yet, from a marketing perspective, it would be difficult to sell vehicles that are intrinsically programmed to sacrifice its occupants under certain conditions, however rare.

The Ethical Robot (Robo-Conscience)

Self-driving cars share a common node with warfare and other high-stakes or lethal applications: the requirement to ensure the system behaves ethically. There are two ways this can be conceptualized. The first is to program specific behaviours which meet the chosen ethical standard – such as, "children must not be harmed". The ethical debate posits a second option: the notion of designing an AI that actually possesses an ethical conscience, able to determine an appropriate action in the absence of a rigid rule.[89] Some argue the machine could behave more conscionably than a human and even experience guilt, while others contend the notion is science fiction, possible only with AGI, and convolutes an already complex ethical debate.[90]

---

[87] H. Joel Trussell, "Why a Special Issue on Machine Ethics," *Proceedings of the IEEE* 106, no. 10 (October 2018): 1774, https://doi.org/10.1109/JPROC.2018.2868336.
[88] Trussell, 1775.
[89] Etzioni and Etzioni, "Incorporating Ethics into Artificial Intelligence," 406.
[90] Lucas, *Engineering, Ethics, and Industry*, 217.

INTERPRETATION AND CONCLUSION

The ethical and policy debate concerning AI is broad, owing to the speed with which the technology has flourished in recent years, and the wide range of its potential use cases. The ethical problem-space for AI has key nodes that address its technological particularities, societal considerations, and several categories of higher-stakes use cases. This paper has presented a framework for describing the ethical problem-space in these terms, with an aim of providing some clarity in a complex and conflated policy debate.

The framework is best interpreted as a system map of the key considerations. Any given policy application is likely to incorporate considerations from several nodes, depending on the proposed implementation. The nodes presented here are not intended to be restrictive or exclusive, but rather to separate the issues into those areas where they are most prevalent. For instance, machine-learning considerations remain relevant in any application where this technology is used. Conversely, a purely conventional AI used in LAWS may be bereft of unique technological considerations, yet still possess substantial ethical weight by nature of being an AS used for warfare, and potentially incorporating dual-use technology licensed by a corporate entity.

It is hoped that the proposed model facilitates the development of robust policies that address the particularities of AI technology without distilling or conflating its key elements. Understanding the ethical landscape is but a simple first step in the monumental task of successfully governing the development and use of AI through policy and law.

BIBLIOGRAPHY

Abrassart, Christophe, Yoshua Bengio, Guillaume Chicoisne, Nathalie de Marcellis-Warin, Marc-Aountoine Dilhac, and others. “Montréal Declaration for a Responsible Development of Artificial Intelligence,” 2018. https://5dcfa4bd-f73a-4de5-94d8-c010ee777609.filesusr.com/ugd/ebc3a3_506ea08298cd4f8196635545a16b071d.pdf.

Akhtar, Naveed, and Ajmal Mian. “Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey.” *IEEE Access* 6 (2018): 14410–30. https://doi.org/10.1109/ACCESS.2018.2807385.

Alom, Md Zahangir, Tarek M. Taha, Christopher Yakopcic, Stefan Westberg, Paheding Sidike, Mst Shamima Nasrin, Brian C. Van Esesn, Abdul A. S. Awwal, and Vijayan K. Asari. “The History Began from AlexNet: A Comprehensive Survey on Deep Learning Approaches.” *ArXiv:1803.01164 [Cs]*, September 12, 2018. http://arxiv.org/abs/1803.01164.

Amnesty International. “The Toronto Declaration: Protecting the Rights to Equality and Non-Discrimination in Machine Learning Systems.” *Access Now* (blog), May 16, 2018. https://www.accessnow.org/the-toronto-declaration-protecting-the-rights-to-equality-and-non-discrimination-in-machine-learning-systems/.

Arkin, Ronald C. “Ethics and Autonomous Systems: Perils and Promises [Point of View].” *Proceedings of the IEEE* 104, no. 10 (October 2016): 1779–81. https://doi.org/10.1109/JPROC.2016.2601162.

Braga, Adriana, and Robert K. Logan. “AI and the Singularity: A Fallacy or a Great Opportunity?” *Information* 10, no. 2 (February 1, 2019): 73. https://doi.org/10.3390/info10020073.

Brykman, Steven. “Douglas Adams Was Right: ‘Genuine People Personalities’ Are Coming to Our Gadgets.” Ars Technica, December 22, 2018. https://arstechnica.com/gadgets/2018/12/douglas-adams-was-right-get-ready-to-talk-with-digital-personalities/.

Char, Danton S., Nigam H. Shah, and David Magnus. “Implementing Machine Learning in Health Care -- Addressing Ethical Challenges.” *The New England Journal of Medicine; Boston* 378, no. 11 (March 15, 2018): 981–83. http://dx.doi.org.cfc.idm.oclc.org/10.1056/NEJMp1714229.

Department of Defense. “Autonomy in Weapon Systems.” Directive. 3000.09, May 8, 2017. https://www.esd.whs.mil/Portals/54/Documents/DD/issuances/dodd/300009p.pdf.

Etzioni, Amitai, and Oren Etzioni. “Incorporating Ethics into Artificial Intelligence.” *The Journal of Ethics; Dordrecht* 21, no. 4 (December 2017): 403–18. http://dx.doi.org/10.1007/s10892-017-9252-2.

“Facebook Increasingly Reliant on A.I. To Predict Suicide Risk.” *All Things Considered*. NPR, November 17, 2018. https://www.npr.org/2018/11/17/668408122/facebook-increasingly-reliant-on-a-i-to-predict-suicide-risk.

Ferretti, Luca, Chris Wymant, Michelle Kendall, Lele Zhao, Anel Nurtay, Lucie Abeler-Dörner, Michael Parker, David Bonsall, and Christophe Fraser. “Quantifying SARS-CoV-2 Transmission Suggests Epidemic Control with Digital Contact Tracing.” *Science*, March 31, 2020. https://doi.org/10.1126/science.abb6936.

Geist, Edward Moore. “It’s Already Too Late to Stop the AI Arms Race—We Must Manage It Instead.” *Bulletin of the Atomic Scientists* 72, no. 5 (September 2, 2016): 318–21. https://doi.org/10.1080/00963402.2016.1216672.

Gunning, David, and David W. Aha. “DARPA’s Explainable Artificial Intelligence Program.” *AI Magazine; La Canada* 40, no. 2 (Summer 2019): 44–58.

Hurley, J. S. “Enabling Successful Artificial Intelligence Implementation in the Department of Defense.” *Journal of Information Warfare; Yorktown* 17, no. 2 (2018): 65–4.

Isaak, Jim, and Mina J Hanna. “User Data Privacy: Facebook, Cambridge Analytica, and Privacy Protection - IEEE Journals & Magazine.” *IEEE Computer* 51, no. 8 (August 2018). https://doi.org/10.1109/MC.2018.3191268.

Johnson, James. “Artificial Intelligence & Future Warfare: Implications for International Security.” *Defense & Security Analysis* 35, no. 2 (April 3, 2019): 147–69. https://doi.org/10.1080/14751798.2019.1600800.

Lachow, Irving. “The Upside and Downside of Swarming Drones.” *Bulletin of the Atomic Scientists* 73, no. 2 (March 4, 2017): 96–101. https://doi.org/10.1080/00963402.2017.1290879.

Lucas, George R. *Engineering, Ethics, and Industry: The Moral Challenges of Lethal Autonomy*. Oxford University Press, 2013. https://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780199926121.001.0001/acprof-9780199926121-chapter-10.

Merrill, Jeremy B. “Liberal, Moderate or Conservative? See How Facebook Labels You.” *The New York Times,* August 23, 2016, sec. U.S. https://www.nytimes.com/2016/08/24/us/politics/facebook-ads-politics.html.

Miller, Seumas. “Machine Learning, Ethics and Law.” *Australasian Journal of Information Systems* 23, no. 0 (May 1, 2019). https://doi.org/10.3127/ajis.v23i0.1893.

National Science & Technology Council. “The National Artificial Intelligence Research and Development Strategic Plan: 2019 Update,” June 2019.

Pigeau, Ross, and Carol McCann. “Re-Conceptualizing Command and Control.” *Canadian Military Journal* 3, no. 1 (2002): 12.

Plows, Johnathan. “Conflation in the Artificial Intelligence Ethics Debate.” Course Paper, Canadian Forces College, 2020.

Russell, Stuart. “Ethics of Artificial Intelligence.” *Nature; London* 521, no. 7553 (May 28, 2015): 415–18.

Sarangi, Saswat, and Pankaj Sharma. *Artificial Intelligence: Evolution, Ethics and Public Policy*. Milton, UNITED KINGDOM: Taylor & Francis Group, 2018. http://ebookcentral.proquest.com/lib/cfvlibrary-ebooks/detail.action?docID=5517469.

Scharre, Paul. *Army of None: Autonomous Weapons and the Future of War*. New York ; London: WW Norton, 2018.

Szegedy, Christian, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. “Intriguing Properties of Neural Networks.” *ArXiv:1312.6199 [Cs]*, February 19, 2014. http://arxiv.org/abs/1312.6199.

“The Campaign To Stop Killer Robots.” Accessed November 20, 2019. https://www.stopkillerrobots.org/about/.

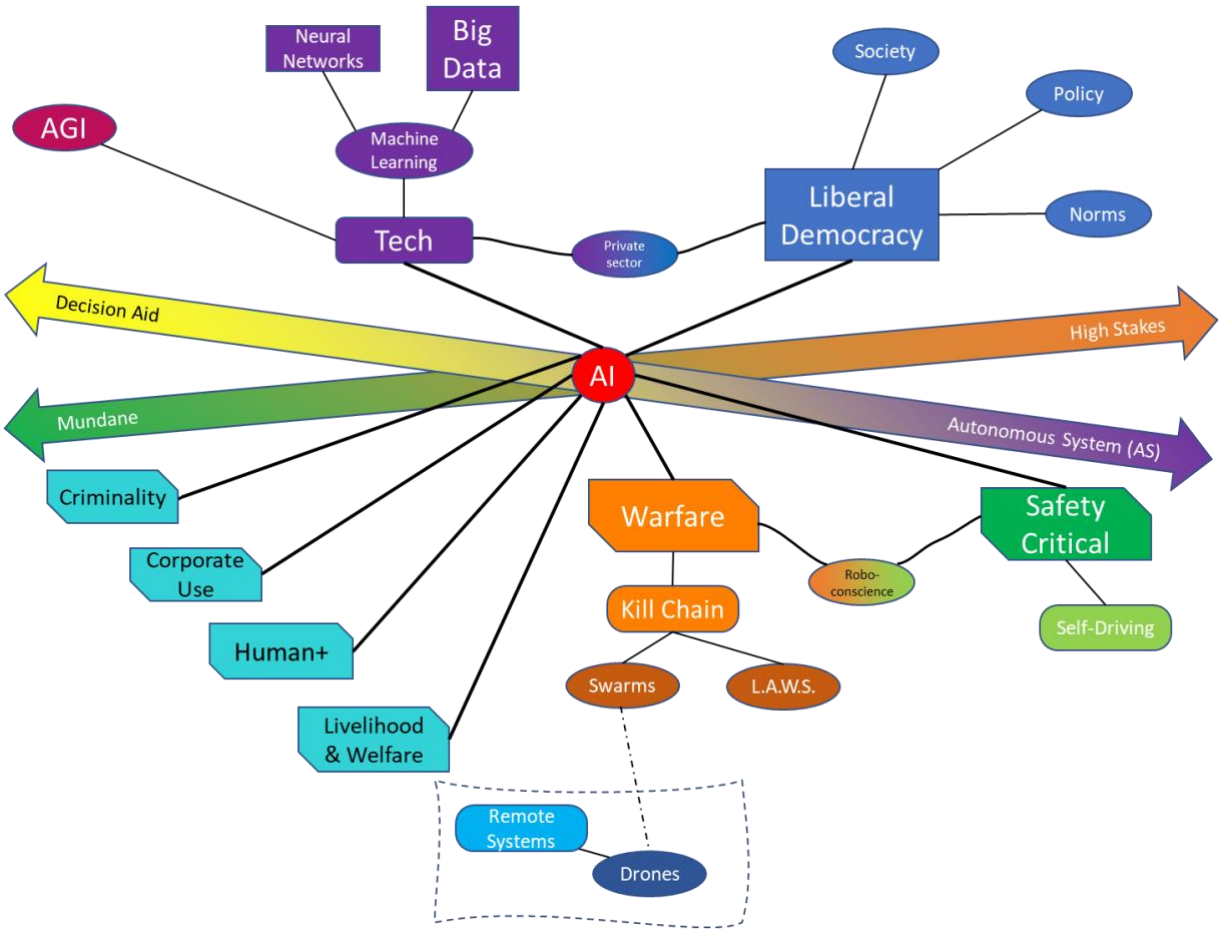
Treasury Board Secretariat. “Directive on Automated Decision-Making.” Government of Canada, February 5, 2019. https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592#appA.

Trussell, H. Joel. “Why a Special Issue on Machine Ethics.” *Proceedings of the IEEE* 106, no. 10 (October 2018): 1774–76. https://doi.org/10.1109/JPROC.2018.2868336.

Wang, Pei. “On Defining Artificial Intelligence.” *Journal of Artificial General Intelligence; Vienna* 10, no. 2 (2019): 1–37. http://dx.doi.org.cfc.idm.oclc.org/10.2478/jagi-2019-0002.

Wasilow, Sherry, and Joelle B. Thorpe. “Artificial Intelligence, Robotics, Ethics, and the Military: A Canadian Perspective.” *AI Magazine; La Canada* 40, no. 1 (Spring 2019): 37–48.

Appendix – The Artificial Intelligence Ethical Problem-Space Map



Notes:

1) Colours and shapes in this diagram are intended to enhance legibility, but do not necessarily correlate with any specific properties or associations.

2) The location of a use-case in the lower portion is not indicative of its position on either of the use-case spectra depicted across the center. As mentioned in the text, most of the use cases discussed live near the "High-Stakes" end of that particular spectrum, but will have potential applications across the entirety of the "Decision Aid – Autonomous System" spectrum.