Artificial Intelligence in Weapons
The Moral Imperative for Minimally-Just Autonomy

Jai Galliott and Jason Scholz

Disclaimer: The views and opinions expressed or implied in the Journal are those of the authors and should not be construed as carrying the official sanction of the Department of Defense, Air Force, Air Education and Training Command, Air University, or other agencies or departments of the US government. This article may be reproduced in whole or in part without permission. If it is reproduced, the Journal of Indo-Pacific Affairs requests a courtesy line.

For military power to be lawful and morally just, future autonomous artificial intelligence (AI) systems must not commit humanitarian errors or acts of fratricide. To achieve this, a preventative form of minimally-just autonomy using artificial intelligence (MinAI) to avert attacks on protected symbols, sites, and signals of surrender is required. MinAI compares favorably to other maximally-just forms proposed to date. This article will examine how fears of speculative AI have distracted from making current weapons more compliant with international humanitarian law. Of particular focus is the Protocol Additional to the Geneva Conventions of 12 August 1949, Article 36.[1] Critics of our approach may argue that machine learning can be fooled, that combatants can commit perfidy to protect themselves, and so forth. This article confronts this issue, including recent research on the subversion of AI, and concludes that the moral imperative for MinAI in weapons remains undiminished.

Introduction

As part of the Campaign to Stop Killer Robots, popular actors, famous business leaders, prominent scientists, lawyers, and humanitarians have called for a ban on autonomous weapons.[2] On 2 November 2017, the campaign sent a letter to Australia's prime minister, Malcolm Turnbull, stating, "Australia's AI research community is calling on you and your government to make Australia the 20th country in the world to take a firm global stand against weaponizing AI." Fearing inaction, these advocates pointed out that the development of autonomous weapons systems would have dire ramifications: "The deadly consequence of this is that machines—*not people*—will determine who lives and dies."[3] It appears that they advocate a complete ban on AI in weapons—an interpretation consistent with their future vision of a world inundated with miniature "slaughterbots."[4]

A ban on AI in weapons may prevent the development of solutions to current humanitarian crises. Every day, the international news media reports incidents with conventional weapons. Consider situations like the following: a handgun stolen from a police officer is subsequently used to kill innocent persons, rifles are used for mass shootings in US schools, vehicles are employed to mow down pedestrians in public places, bombs are deployed to strike religious sites, a guided-bomb is used to strike a train bridge as an unsuspecting passenger train passes, a missile is fired to strike a Red Cross facility, and so forth. With the development of AI weapons, preventing these types of incidents might be possible. These are real situations where an autonomous weapon system equipped with AI might intervene to save lives.

Confusion about the means needed to achieve a desired state of nonviolence is not new. A general disdain for simple technological solutions aimed at a better state of peace was prevalent in the antinuclear campaign—spanning the whole confrontation period with the Soviet Union and recently renewed with the invention of miniaturized warheads and the campaign to ban land mines in the late nineties.[5] It does not seem unreasonable to ask why weapons with advanced seekers could not embed AI to identify a symbol of the Red Cross and abort an ordered strike. Additionally, the location of protected sites of religious significance, schools, and hospitals could be programmed into weapons to constrain their actions. Preventing weapons from firing at humans by an unauthorized user could also be specified. Why should we not begin to test such innovations so that they might be ensconced in international weapons review standards?

This article asserts that autonomous systems are not likely to be capable of action leading to the attribution of moral responsibility in the near term. However, these systems might today autonomously execute value-laden decisions embedded in their design and code, so they can perform actions to meet enhanced ethical and legal standards.[6]

The Ethical Machine Spectrum

A distinction between the two ends of the spectrum of ethical capability needs to be made. A maximally-just autonomy using artificial intelligence (MaxAI) guided by acceptable and nonacceptable actions has the benefit of ensuring ethically obligatory lethal action—even when system engineers of a subordinate system may not have recognized the need or possibility of the relevant lethal action.

However, a maximally-just ethical robot requires extensive ethical engineering. Ronald Arkin's *ethical governor* represents the most advanced prototype effort toward a maximally-just system.[7] The ethical governor provides an assessment on proposed lethal actions consistent with the laws of war and the rules of engagement. The maximally-just position is apparent from the explanation of the operation of the constraint interpreter, which is a key part of the governor: "The constraint application process is responsible for reasoning about the active ethical constraints and ensuring that the resulting behavior of the robot is ethically permissible."[8] The constraint system—based on complex deontic and predicate logic—evaluates the proposed actions generated by the tactical reasoning engine of the system based on an equally complex data structure. Reasoning about the full scope of what is ethically permissible—including notions of proportionality and rules of engagement as Arkin describes—is prone to difficulty.

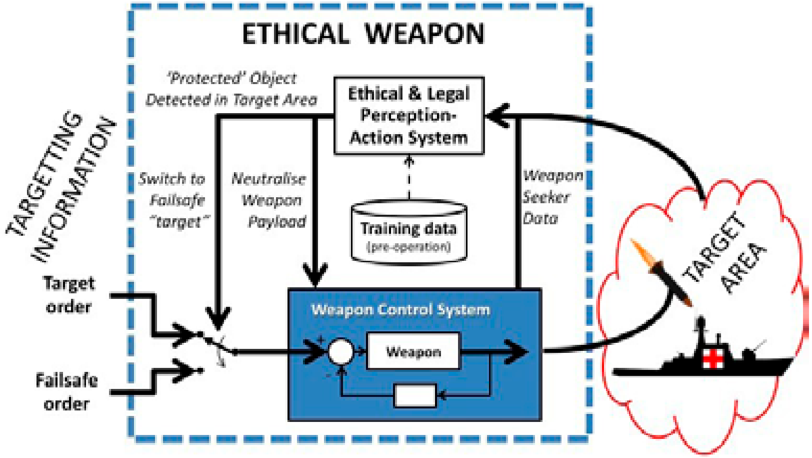
ETHICAL WEAPON
'Protected' Object Detected in Target Area
Ethical & Legal Perception-Action System
TARGETTING INFORMATION
Switch to Failsafe "target"
Neutralise Weapon Payload
Training data (pre-operation)
Weapon Seeker Data
TARGET AREA
Target order
Failsafe order
Weapon Control System
+
-
Weapon

Figure 1. A MinAI ethical weapon. Such a weapon has the ability to disobey a target order in favor of a failsafe specification if an unexpected legally- or ethically-protected object or behavior is perceived in the effected target area. Target data is sourced externally to the weapon.

In contrast, a MinAI ethical robot, while still a constraint-driven system, could operate without a proper ethical governor, possessing only an elementary suppressor of human-generated lethal action that would activate in accordance with a much narrower set of constraints (hard-coded rather than soft-coded)—meaning less system interpretation would be required. MinAI deals with what is *ethically impermissible*, basing constraints on the need to identify and avoid protected objects and behaviors. Specifically avoided are lawfully protected symbols and locations, signs of surrender (including beacons), and sites that are hors de combat. It is important to note that these AI constraints range in scale of difficulty and will continue to improve as AI technologies advance. The conceptual model for a MinAI ethical weapon is illustrated in figure 1.

While MinAI will be more limited in a technical nature, it may be more morally desirable in outcomes than MaxAI in a range of specific circumstances. The former will not take active lethal or nonlethal action against protected persons or infrastructure. In contrast, MaxAI involves the codification of normative values into rule sets and the interpretation of a wide range of inputs through the application of complex and potentially imperfect machine logic. This more-complex algorithmic morality—while potentially desirable in some circumstances—involves a greater possibility of actively introducing fatal errors, particularly in terms of managing conflicts between interests.

Cognizant of the above dilemma, this article suggests that to meet fundamental moral obligations to humanity, we are ethically justified to develop MinAI systems. The ethical agency embedded in the machine and, thus, technologically mediated by the design, engineering, and operational environment, is less removed from the human moral agency than it is in a MaxAI system. MaxAI development is supererogatory in the sense that it may be morally beneficial in particular circumstances but is not necessarily morally required—and may even be demonstrated to be unethical.

Minimally-Just AI as "Hedging One's Bets"

To the distaste of some, one might argue that the moral desirability of MinAI will decrease in the near future as the AI underpinning MaxAI becomes more robust and as we move away from rule-based and basic neural network systems toward artificial general intelligence (AGI). Furthermore, the argument is that resources should be dedicated to the development of maximal ethical robots. To be clear, there have been a number of algorithm success stories announced in recent years, across all the cognate disciplines. The ongoing development of algorithms as a basis for the success of AlphaGo and LibratusMuch has garnered much attention.[9] These systems compete against the best human Go and Poker players, winning against players who have made acquiring deep knowledge of these games their life's work. The result of these preliminary successes has been a dramatic increase in media reporting and interest in the potential opportunities and pitfalls associated with the development of AI. Not all of these reports are accurate, and some have negatively impacted public perception of AI, fueling the kind of dystopian visions advanced by the Campaign to Stop Killer Robots, as mentioned earlier.

The speculation that superintelligence is on the foreseeable horizon—with AGI realization timelines predicted in 20–30 years—reflects the success stories, while omitting discussions of the recent failures in AI. Many failures are unreported due to commercial and classification reasons. One example is Microsoft's Tay AI Bot, a machine learning chatbot that learns from interactions with digital users. After a short period of operation, Tay developed an ego or character that was strongly sexual and racialized, and ultimately Microsoft had to withdraw the bot from service.[10] Facebook had similar problems with its AI message chatbots assuming undesirable characteristics.[11] Additionally, a number of autonomous road vehicles have been involved in motor vehicle accidents where the relevant systems were incapable of handling the scenario and quality-assurance practices failed to factor for such events.

There are currently irresolvable problems with the complex neural networks on which the successes in AI are based. These bottom-up systems can learn well in controlled environments and easily outperform humans in these scenarios based on data structures and their correlations, but these systems cannot match the top-down rationalizing power of human beings in more open environments, such as road systems and conflict zones. Such systems are risky in these environments because they require strict compliance with laws and regulations. It would be difficult to question, interpret, explain, supervise, and control these systems because deep-learning systems cannot easily track their own "reasoning."[12]

Just as importantly, when more-intuitive and, therefore, less-explainable systems come into wide operation, it may not be so easy to revert to earlier-stage systems, as human operators become reliant on the system to make difficult decisions. The danger becomes that operators' own moral decision-making skills may have deteriorated over time.[13] In the event of failure, total system collapse could occur, with devastating consequences if such systems were committed to a mission-critical operation required in armed conflict.

There are, moreover, issues associated with functional complexity and the practical computational limits imposed on mobile systems that need to be capable of independent operation in the event of a communications failure. The computers required for AGI-level systems may not be subject to miniaturization or simply may not be sufficiently powerful or cost-effective for the intended purpose, especially in a military context in which autonomous weapons are sometimes considered disposable platforms.[14] The hope for advocates of AGI is that computer processing power and other system components will continue to become dramatically smaller, cheaper, and powerful, but there is no guarantee that Moore's law, which supports such expectations, will continue to reign true without extensive progress in the field of quantum computing.

MaxAI at this point in time, whether or not AGI should eventuate, appears a distant goal to deliver a potential result that unguaranteed. A MinAI system, on the other hand, seeks to ensure that the obvious and uncontroversial benefits of AI are harnessed, while the associated risks are kept under control by normal military targeting processes. Decision makers need to take action now to stave off grandiose visions that may not eventuate and instead deliver a positive result with technology that already exists.

Implementation

International Humanitarian Law Article 36 states, "In the study, development, acquisition or adoption of a new weapon, means or method of warfare, a High Contracting Party is under an obligation to determine whether its employment would, in some or all circumstances, be prohibited by this Protocol or by any other rule of international law applicable to the High Contracting Party."[15] The Commentary of 1987 to the Article further indicates that a state must review not only new weapons but also any existing weapon that is modified in a way that alters its function—or a weapon that has already passed a legal review that is subsequently modified.[16] Thus, the insertion of MinAI in a weapon would require Article 36 review.

The customary approach to assessment to comply with Article 36 covers the technical description and performance of the weapon and assumes humans assess and decide weapons use.[17] AI poses challenges for assessment under Article 36 in situations where there was once human-decision functions were clearly separated from weapon technical function assessment. Assessment approaches need to extend to embedded decision making and acting capability for MinAI.

Although Article 36 deliberately avoids imposing how such a determination made, it might be in the interests of the International Committee of the Red Cross—and humanity as a whole—to do so in this specific case. Consider the first reference in international treaties to the need to conduct legal reviews of new weapons.[18] As a precursor to Article 36, the Saint Petersburg Declaration has a broader scope: "The Contracting or Acceding Parties reserve to themselves to come hereafter to an understanding whenever a precise proposition shall be drawn up in view of future improvements which science may effect in the armament of troops, in order to maintain the principles which they have established, and to conciliate the necessities of war with the laws of humanity."[19] MinAI in weapons and autonomous systems is precisely such a proposition. The ability to improve humanitarian outcomes through embedded weapon capability to identify and prevent attack on protected objects might form a recommended standard.

The sharing of technical data and algorithms for achieving this standard through Article 36 would drive down the cost of implementation and expose systems to countermeasures that improve their security.

Humanitarian Counter-Countermeasures

Critics may argue that combatants will develop countermeasures aimed at spoiling the intended humanitarian effects of MinAI in weapons and autonomous systems. However, it is antihumanitarian and potentially illegal to field countermeasures to MinAI. Yet, many actors do not comply with the rule of law. Therefore, it is necessary to consider countermeasures to MinAI that may seek to degrade, damage, destroy, or deceive the capability so as to secure the targeted system.

Degradation, Damage, or Destruction

It is expected that lawfully targeted enemies will attempt to destroy or degrade weapon performance to prevent MinAI from achieving its intended mission. Such countermeasures could include attack against the weapon seeker or other means. Such an attack may degrade, damage, or destroy the MinAI capability. If the act is in self-defense, this is not a behavior expected of a humanitarian object and, thus, the function of the MinAI is not required anyway.

If the degradation, damage, or destruction is targeted against the MinAI in order to cause a humanitarian disaster, it would be a criminal act. However, for this to occur, the legal status of the target would have had to have been neglected as a precursor, prior to this act, which ought to be the primary cause for concern.

Deception

Combatants might simply seek to deceive the MinAI capability by using something akin to a Red Cross or Red Crescent symbol to protect themselves, thereby averting an otherwise lawful attack. This is an act of perfidy covered under IHL Article 37. Yet, such an act may serve to improve distinction, by crosschecking perfidious sites with the Red Cross to identify anomalies. Furthermore, the Red Cross symbol is an distinctive marker, so wide-area surveillance might be sensitive to subsequent attempts at such deception. Further, it is for this reason that we distinguish that MinAI ethical weapons respond only to the *unexpected* presence of a protected object or behavior. This response is a decision made in the targeting process and is external to the ethical weapon, as illustrated in figure 1. A log for accountability and subsequent review of the action will be generated.

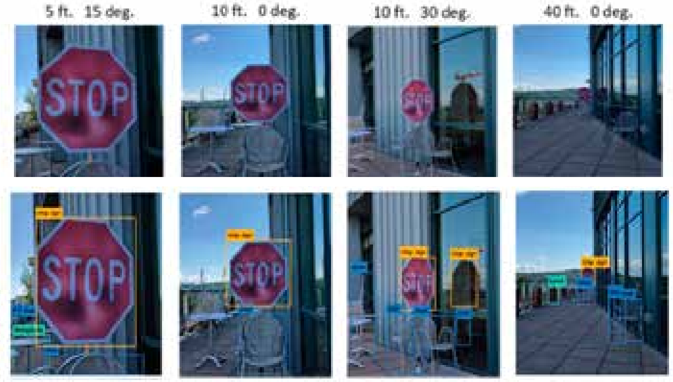

Figure 2. (Top) Adversarial 2D camouflage to a stop sign imitating wear and tear, using a convolutional neural network—a class of deep, feed-forward artificial neural networks, most commonly applied to analyzing visual imagery—on the Laboratory for Intelligent and Safe Automobiles road-signs database, achieves 100-percent success classifying each of these as 45-mph-speed signs. (Kevin Eykholt et al., "Robust Physical-World Attacks on Deep Learning Visual Classification" (paper, 2018 Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, 18–22 June 2018), https://arxiv.org/pdf/1707.08945.pdf.) (Bottom) For a detector followed by a classifier—a mapping from unlabeled instances to discrete classes—achieves 100-percent failure, correctly identifying these as stop signs every time. (Jiajun Lu et al., "Standard Detectors Aren't (Currently) Fooled by Physical Adversarial Stop Signs," 26 October 2017, https://arxiv.org/pdf/1710.03337.pdf.)

The highest-performing object-recognition systems are neural networks; yet, the high dimensionality that gives them that performance level may in itself be a vulnerability. Researchers have discovered a phenomenon related to stability given small perturbations to inputs, where a non-random perturbation imperceptible to humans could be applied to a test image and result in an arbitrary change to its estimate.[20] A significant body of work has since emerged on these "adversarial examples."[21] Of the many and varied forms of attack, there exists a range of countermeasures. A subclass of adversarial examples of relevance to MinAI are those that can be applied to two- and three-dimensional physical objects to change their appearance to the machine. Recently adversarial algorithms have been used to generate *camouflage paint* and even 3-D printed objects resulting in errors for standard deep network classifiers.[22] Concerns include the possibility to paint a Red Cross symbol on an object that is recognizable by a weapon seeker yet invisible to humans and the dual case illustrated in figure 2 of painting over a protection symbol with marking resembling weathered patterns unnoticeable to humans yet resulting in an algorithm being rendered unable to recognize the sign—in this case a traffic stop sign symbol, which is of course similar to a Red Cross symbol.

In contrast to these results popularized by online media, researchers have demonstrated *no errors* on the same experimental setup as the stop-sign scenario above and in live trials. These researchers explained that the original team had confused *detectors*—a class of deep, feed-forward artificial neural networks, most commonly applied to analyzing visual imagery—(like Faster region-based convolutional neural networks [R-CNN]) with *classifiers*—a mapping from unlabeled instances to discrete classes.[23] Methods used in the first of these experiments appear to be at fault due to pipeline problems, including perfect manual cropping (a proxy for a detector that has been assumed away) and rescaling before applying to a classifier. Outside of the lab environment, it remains difficult to conceive of a universal defeat for a detector under various real-world angle, range, and light conditions, but further research is required.

Global open access to MinAI code and data, for example Red Cross imagery and video scenes in "the wild," would have the significant advantage of ensuring these techniques continue to be tested and hardened under realistic conditions and architectures. Global access to MinAI algorithms and data sets would speed implementation, offering low-cost solutions for nations that might not otherwise afford such innovations, and exert moral pressure on defense companies that do not use this resource.

International protections against countermeasures targeting MinAI might be mandated. If such protections were accepted, it would strengthen the case for the employment of MinAI, but in the absence of such protections, the moral imperative for MinAI in weapons remains undiminished in light of countermeasures.

Conclusion

This article presented a case MinAI that could make life-saving decisions in the world today. The hope is that the significant resources spent on reacting to speculative fears of campaigners might one day be spent mitigating the suffering of people caused by weapons that lack MinAI. JIPA

Notes

[1] *See* Article 36. International Committee of the Red Cross, "Protocol Additional to the Geneva Conventions of 12 August 1949, and Relating to the Protection of Victims of International Armed Conflicts (Protocol I), 8 June 1977," accessed 24 October 2018, https://ihl-databases.icrc.org/applic/ihl/ihl.nsf/7c4d08d9b287a42141256739003e636b/f6c8b9fee14a77fdc125641e0052b079.

[2] "The Solution," *Campaign to Stop Killer Robots*, 2018, https://www.stopkillerrobots.org/the-solution/.

[3] Damminda Alahakoon et al, "RE: An International Ban on the Weaponization of Artificial Intelligence (AI)," website of Prof. Toby Walsh, University of New South Wales, 2 November 2017, https://www.cse.unsw.edu.au/~tw/letter.pdf.

[4] "Keep Killer Robots Science Fiction," *Ban Lethal Autonomous Weapons* (vlog), 10 November 2017, https://autonomousweapons.org/slaughterbots/.

[5] The United States, of course, never ratified the Ottawa treaty but rather chose a technological solution to end the use of persistent land mines—land mines that can be set to self-destruct or deactivate after a predefined time period—making them considerably less problematic when used in clearly demarcated and confined zones such as the Korean Demilitarised Zone. For information *see* Lorraine Boissoneault, "The Historic Innovation of Land Mines—And Why We've Struggled to Get Rid of Them," *Smithsonian*, accessed October 24, 2018, https://www.smithsonianmag.com/innovation/historic-innovation-land-minesand-why-weve-struggled-get-rid-them-180962276/.

[6] Patrick Chisan Hew, "Artificial Moral Agents Are Infeasible with Foreseeable Technologies," *Ethics and Information Technology; Dordrecht* 16, no. 3 (September 2014): 197–206, https://doi.org/10.1007/s10676-014-9345-6.

[7] Ronald C. Arkin, Patrick Ulam, and Brittany Duncan, "An Ethical Governor for Constraining Lethal Action in an Autonomous System" (technical report, Fort Belvoir, VA: Defense Technical Information Center, 1 January 2009), https://doi.org/10.21236/ADA493563.

[8] Ibid.

[9] James O'Malley "The 10 Most Important Breakthroughs in Artificial Intelligence," *TechRadar*, a10 January 2018, https://www.techradar.com/news/the-10-most-important-breakthroughs-in-artificial-intelligence.

[10] John West, "Microsoft's Disastrous Tay Experiment Shows the Hidden Dangers of AI," *Quartz*, 2 April 2016, https://qz.com/653084/microsofts-disastrous-tay-experiment-shows-the-hidden-dangers-of-ai/.

[11] Alex Hern, "Please, Facebook, Don't Make Me Speak to Your Awful Chatbots," *Guardian*, 29 April 2016, https://www.theguardian.com/technology/2016/apr/29/please-facebook-dont-make-me-speak-to-your-awful-chatbots.

[12] Martin Ciupa, "Is AI in Jeopardy? The Need to Under Promise and Over Deliver—The Case for Really Useful Machine Learning," in *Computer Science & Information Technology (CS & IT)* (Fourth International Conference on Computer Science and Information Technology, Academy & Industry Research Collaboration Center (AIRCC), 2017), 59–70, https://doi.org/10.5121/csit.2017.70407.

[13] Jai Galliott, "The Limits of Robotic Solutions to Human Challenges in the Land Domain," *Defence Studies* 17, no. 4 (2 October 2017): 327–45, https://doi.org/10.1080/14702436.2017.1333890.

[14] Ibid.

[15] Protocol Additional to the Geneva Conventions of 12 August 1949, and relating to the Protection of Victims of International Armed Conflicts (Protocol I), 8 June 1977, https://ihl-databases.icrc.org/ihl/WebART/470-750045?OpenDocument.

[16] Protocol Additional to the Geneva Conventions of 12 August 1949, and relating to the Protection of Victims of International Armed Conflicts (Protocol I), 8 June 1977, Commentary of 1987, "New Weapons," 421, https://ihl-databases.icrc.org/applic/ihl/ihl.nsf/Comment.xsp?action=openDocument&documentId=F095453E41336B76C12563CD00432AA1.

[17] "A Guide to the Legal Review of New Weapons, Means and Methods of Warfare: Measures to Implement Article 36 of Additional Protocol I of 1977," *International Review of the Red Cross* 88, no. 864 (December 2006): 931, https://doi.org/10.1017/S1816383107000938.

18. “Declaration Renouncing the Use, in Time of War, of Certain Explosive Projectiles. Saint Petersburg, 29 November/11 December 1868,” accessed 24 October 2018, http://www.gwpda.org/1914m/gene68.html.

19. Ibid.

20. Christian Szegedy et al., “Intriguing Properties of Neural Networks,” *Computing Research Repository*, 19 February 2014, https://arxiv.org/pdf/1312.6199.pdf.

21. Naveed Akhtar and Ajmal Mian, “Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey,” *IEEE Access* 6 (2 January 2018), 14410–30, https://arxiv.org/pdf/1801.00553.pdf.

22. Kevin Eykholt et al., “Robust Physical-World Attacks on Deep Learning Visual Classification” (paper, 2018 Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, 18–22 June 2018), https://arxiv.org/pdf/1707.08945.pdf.

23. Jiajun Lu et al., “Standard Detectors Aren’t (Currently) Fooled by Physical Adversarial Stop Signs,” 26 October 2017, https://arxiv.org/pdf/1710.03337.pdf

Dr. Jai Galliott

Dr. Jai Galliott leads the Values in Defence & Security Technology Group within the University of New South Wales at the Australian Defence Force Academy. As a former Royal Australian Navy officer and Australian Army Research Fellow, his recent books include Military Robots: Mapping the Moral Landscape; Ethics and the Future of Spying: Technology, National Security and Intelligence Collection; Super Soldiers: The Ethical, Legal and Social Implications; Commercial Space Exploration: Ethics Policy and Governance; and Force Short of War in Modern Conflict: Jus ad Vim. He is nonresident fellow at the Modern War Institute at West Point and visiting fellow with the Centre for Technology and Global Affairs at the University of Oxford.

Professor Jason Scholz

Professor Jason Scholz holds a bachelors degree in electronic engineering from the University of South Australia and PhD in electrical engineering from the University of Adelaide. He has over 75 refereed open publications and patents in telecommunications, signal processing, artificial intelligence, and human decision-making according to Google Scholar--with several hundred citations. He is chief scientist for the Trusted Autonomous Systems Defence Cooperative Research Centre, a nonprofit company advancing industry-led, game-changing projects for the Australian Department of Defence. He also works for Defence Science and Technology, leading the strategic research initiative on trusted autonomous systems. This involves research, development, and showcasing of high-impact technologies for persistent autonomy, machine cognition, and human-machine integration in close partnership with overseas governments, academia, and industry to deliver game-changing impact for Australian defense and national security. He is an assessor for the Australian Research Council and graduate of the Australian Institute of Company Directors. His professorial position is adjunct with the Australian Defence Force Academy at the University of New South Wales.