AI IN WEAPONS: THE MORAL IMPERATIVE FOR MINIMALLY-JUST AUTONOMY

Prof. Jason Scholz[1]
Chief Scientist & Engineer
Trusted Autonomous Systems Defence CRC
Queensland, Australia
jason.scholz@tasdcrc.com.au

Dr. Jai Galliott
Group Leader - Values in Defence & Security Technology
University of New South Wales @ ADFA
Canberra, Australian Capital Territory, Australia
j.galliott@unsw.edu.au

***Abstract*— For land power to be lawful and morally just, future autonomous systems must not commit humanitarian errors or acts of fratricide. To achieve this, we distinguish a novel preventative form of minimally-just autonomy using artificial intelligence (MinAI) to avert attacks on protected symbols, protected sites and signals of surrender. MinAI compares favourably with respect to maximally-just forms proposed to date. We examine how fears of speculative artificial general intelligence has distracted resources from making current weapons more compliant with international humanitarian law, particularly Additional Protocol 1 of the Geneva Convention and its Article 36. Critics of our approach may argue that machine learning can be fooled, that combatants can commit perfidy to protect themselves, and so on. We confront this issue, including recent research on the subversion of AI, and conclude that the moral imperative for MinAI in weapons remains undiminished.**

Keywords—machine ethics; Reasoning and Learning Machines; autonomous weapon systems; legal robots.

I. INTRODUCTION

Popular actors, famous business leaders, prominent scientists, lawyers and humanitarians, as part of the Campaign to Stop Killer Robots, have called for a ban on autonomous weapons. On 2 November 2017, a letter organised by the Campaign was sent to Australia's Prime Minister 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 – a 'consequence of this is that machines—*not people*—will determine who lives and dies' [1]. It appears that they mean a complete ban on **AI in weapons**, an interpretation consistent with their future vision of a world awash with miniature 'slaughterbots' [2].

We hold that a ban on AI in weapons may prevent the development of solutions to current humanitarian crises. Every day in the World News real problems are happening with conventional weapons. Consider situations like: a handgun stolen from a police officer and subsequently used to kill innocent persons, rifles used for mass shootings in US schools, vehicles used to mow down pedestrians in public places, bombing of religious sites, a guided-bomb strike on a train bridge as an unexpected passenger train passes, a missile strike on a Red Cross facility, and so on – all might be prevented. These are real situations where a weapon or autonomous system equipped with AI might intervene to *save lives*.

Confusion about the means to achieve desired nonviolence is not new. A general disdain for simple technological solutions aimed at a better state of peace was prevalent in the anti-nuclear campaign spanning the whole confrontation period with the Soviet Union, recently renewed with the invention of miniaturised warheads, and the campaign to ban land mines in the late nineties.[2] Yet, 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. Or why the location of protected sites of religious significance, schools or hospitals might be programmed into weapons to constrain their actions, or guns prevented from firing by an unauthorised user pointing it at humans. And why initiatives cannot begin to test these innovations wo that they might be ensconced in International weapons review standards?

We assert that while autonomous systems are likely to be incapable of action leading to the attribution of *moral responsibility* [3] in the near term, they might today autonomously execute value-laden decisions embedded in their design and in code, so they can perform actions to meet enhanced ethical and legal standards.

II. THE ETHICAL MACHINE SPECTRUM

Let us discern between two ends of a spectrum of ethical capability. A maximally just 'ethical machine' (MaxAI) guided by both acceptable and non-acceptable actions has the benefit of ensuring that ethically obligatory lethal action is taken, even when system engineers of a lesser system may not have recognised the need or possibility of the relevant lethal action. However, a maximally-just ethical robot requires extensive ethical engineering. Arkin's 'ethical governor' [4] represents probably the most advanced prototype effort towards a maximally-just system. The ethical governor provides assessment on proposed lethal actions consistent with the laws of war and 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

[1] Adjunct position at UNSW @ ADFA.

[2] The United States, of course, never ratified the Ottawa treaty but rather chose a technological solution to end the use of persistent landmines – landmines that cannot 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.

application process is responsible for reasoning about the active ethical constraints and ensuring that the resulting behavior of the robot is ethically permissible' [4]. That is, 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 a hard problem.

In contrast, a MinAI 'ethical robot', while still a constraint driven system, could operate without an 'ethical governor' proper and need only contain an elementary suppressor of human-generated lethal action that would activate in accordance with a much narrower set of constraints that may be hard rather than soft coded, meaning far less system 'interpretation' would be required. MinAI deals with what is *ethically impermissible*. These constraints would be based around the need to identify and avoid 'protected' objects and behaviours. Specifically, lawfully-protected symbols, protected locations, basic signs of surrender (including beacons), and potentially those that are *hors de combat*, noting of course that these AI problems range from easy to more difficult – but not impossible – and will continue to improve with AI technologies. The basic concept for a MinAI Ethical Weapon is illustrated in figure 1.

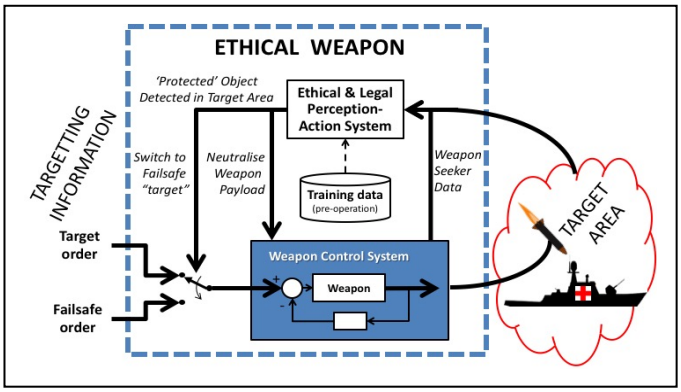

Fig 1. A MinAI Ethical Weapon has the ability to disobey a target order in favour of a failsafe specification if an unexpected legally- or ethically-protected object or behaviour is perceived in the effected target area. Target data is sourced externally to the weapon.

Noteworthy is that while MinAI will always be more limited in technical nature, it may be more morally desirable in that it will yield outcomes that are as good as or possibly even better than MaxAI in a range of specific circumstances. The former will never take active lethal or non-lethal action to harm 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.

Cognisant of the above, our suggestion is that in terms of meeting our fundamental moral obligations to humanity, we are ethically justified to develop MinAI systems. The ethical agency of said system, whilst embedded in the machine and thus technologically mediated by the design, engineering and operational environment, is fewer steps removed from human moral agency than in a MaxAI system. We would suggest that 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.

III. Minimally-Just AI as Hedging One's Bets

To the distaste of some, it might be argued that the moral desirability of MinAI will decrease in the near future as the AI underpinning MaxAI becomes more robust, and we move away from rule-based and basic neural network systems toward artificial general intelligence (AGI), and that resources should therefore 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. Much attention has been given to the ongoing development of Algorithms as a basis for the success of AlphaGo [5] and Libratus. These systems are competing against the best human Go and Poker players and winning against those 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 on, and interest in, the potential opportunities and pitfalls associated with the development of AI, not all of which are accurate and some of which has negatively impacted public perception of AI, fuelling 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 timelines in the realm of 20-30 years, reflects the success stories while omitting discussion of recent failures in AI. Many of these undoubtedly go unreported for commercial and classification reasons, but Microsoft's Tay AI Bot, a machine learning chatbot that learns from interactions with digital users, is but one example. After a short period of operation, Tay developed an 'ego' or 'character' that was strongly sexual and racialized, and ultimately had to be withdrawn from service. Facebook had similar problems with its AI message chatbots assuming undesirable characteristics and a number of autonomous road vehicles have now been involved in motor vehicle accidents where the relevant systems were incapable handling the scenario and quality assurance practices failed to factor for such events.

There are also known and currently irresolvable problems with the complex neural networks on which the successes in AI have mostly been based. These bottom-up systems can learn well in tight domains and easily outperform humans in these scenarios based on data structures and their correlations, but they cannot match the top-down rationalising power of human beings in more open domains such as road systems and conflict zones. Such systems are risky in these environments because they require strict compliance with laws and regulations and it would be difficult to question, interpret, explain, supervise and control them by virtue of the fact that deep learning systems cannot easily track their own 'reasoning' [6].

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, with the danger that their own moral decision-making skills may have deteriorated over time [7]. In the event of failure, total system collapse could occur with devastating consequences if such systems were committed to 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 [6]. 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 is far from guaranteed. A MinAI system, on the other hand, seeks to ensure that the obvious and uncontroversial benefits of artificial intelligence are harnessed while the associated risks are kept under control by normal military targeting processes. Action needs to be taken now to intercept grandiose visions that may not eventuate and instead deliver a positive result with technology that already exists.

IV. Implementation

International Humanitarian Law Article 36 states [8], '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.' 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. Thus, the insertion of minimally-just AI in a weapon would require Article 36 review.

The customary approach to assessment [9] to comply with Article 36 covers the technical description and technical performance of the weapon and assumes humans assess and decide weapon use. Artificial Intelligence poses challenges for assessment under Article 36 where there was once a clear separation of human decision functions 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 is carried out, it might be in the interests of the International Committee of the Red Cross and humanity to do so in this specific case. Consider the first reference in international treaties to the need to carry out legal reviews of new weapons [10]. As a precursor to IHL Article 36 this treaty 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.' MinAI in weapons and autonomous systems is such a precise 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 means through Article 36 would drive down the cost of implementation and expose systems to countermeasures that improve their hardening.

V. Humanitarian Counter-Countermeasures

Critics may argue that combatants will develop countermeasures that aim to spoil the intended humanitarian effects of MinAI in weapons and autonomous systems. We claim it to be anti-humanitarian to field countermeasures to MinAI and potentially illegal to do so. Yet, many actors do not comply with the rule of law. So, it is necessary to consider countermeasures to MinAI that may seek to degrade, damage, destroy, or deceive the capability so as to harden it.

A. Degradation, Damage or Destruction

It is expected that lawfully-targeted enemies will attempt to destroy or degrade weapon performance to prevent it from achieving the intended mission. This could include attack to the weapon seeker or other means. Such an attack may as a consequence degrade, damage or destroy the MinAI capability. If the act is in self Defence, this is not a behavior one would expect from 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 appreciation of the target would have had to have failed as a precursor, prior to this act, which is the primary cause for concern.

B. Deception

Combatants might simply seek to deceive the MinAI capability by using say, a symbol of the Red Cross or Red Crescent 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 cross-checking perfidious sites with the Red Cross to identify anomalies. Further, a Red Cross is an obvious marker, so wide area surveillance might be sensitive to picking up new instances. 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. Of course, this is a decision made in the targeting process (which is external to the ethical weapon) as per Figure 1, and would be logged for accountability and subsequent review of action.

The highest performing object recognition systems are neural networks, yet, the high dimensionality that gives them that performance, may in itself be a vulnerability. Szedgy *et al* [11] 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. A significant body of work has since emerged on these "adversarial examples" [12]. Of the many and varied forms of attack, there also exist 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 [13] adversarial algorithms been used to generate 'camouflage paint' and even 3D printed objects resulting in errors for standard deep network classifiers. Concerns include the possibility to paint a Red Cross symbol on an object that recognizable by a weapon seeker yet invisible to humans, or the dual case illustrated in figure 2 of painting over a protection symbol with marking resembling weathered patterns unnoticeable to humans yet resulting an algorithm unable to recognize the sign (in this case a traffic stop sign symbol, which is of course similar to a Red Cross symbol).

Fig 2. (Top) Adversarial 2D camouflage to a stop sign imitating wear using CNN on the LISA road signs database, achieves 100% success classifying each of these as 45 mph speed signs [13]. (Bottom) For a *detector* followed by classifier achieves 100% failure, correctly identifying these as stop signs every time [14].

In contrast to these results popularized by online media, Lu *et al* [14] demonstrate *no errors* on the same experimental setup as [13] and in live trials, explaining that the authors of [14] have confused *detectors* (like Faster RCNN) with *classifiers*. Methods used in [13] appear to be at fault due to pipeline problems, including perfect manual cropping (a proxy for a detector which has been assumed away) and rescaling before applying to a classifier. In the real world it remains difficult to conceive of a universal defeat for a detector under various real-world angle, range and light conditions, yet 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 ease uptake, especially as low-cost solutions for Nations that might not otherwise afford such innovations, as well as exerting moral pressure on Defence companies that do not use this resource.

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

VI. Conclusion

We have presented a case for autonomy in weapons that could make life-saving decisions in the world today. We hope in future that the significant resources spent on reacting to speculative fears of campaigners might one day be spent mitigating the definitive suffering of people caused by weapons which lack minimally-just autonomy based on artificial intelligence.

References

[1] T. Walsh, https://www.cse.unsw.edu.au/~tw/letter.pdf

[2] Campaign to Stop Killer Robots. Slaughterbots video. See: http://autonomousweapons.org

[3] P.C. Hew, Artificial moral agents are infeasible with foreseeable technologies. Ethics and Information Technology 16 (3):197-206, 2014.

[4] R.C. Arkin, P. Ulam, B. Duncan, "An Ethical Governor for Constraining Lethal Action in an Autonomous System", Technical Report GIT-GVU-09-02.

[5] J. O'Malley "The 10 Most Important Breakthroughs in Artificial Intelligence", Tech Radar, 2018. See: https://www.techradar.com/news/the-10-most-important-breakthroughs-in-artificial-intelligence

[6] M. Ciupa, Is AI In Jeopardy? The Need to Under Promise and Over Deliver – The case for Really Useful Machine Learning in Dhinaharan Nagamalai et. Al. (Eds) 4th Int. Conf. on Computer Science and Information Technology (CoSIT 2017) Geneva, Switzerland, March 25-26, 2017, pp. 59-70.

[7] J. Galliott The limits of robotic solutions to human challenges in the land domain, Defence Studies 17 (4), 2017, 327-345.

[8] Article 36 of Protocol I Additional to the 1949 Geneva Conventions.

[9] International Committee of the Red Cross, 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, Vol. 88 No. 864 Dec. 2006, Geneva. https://www.icrc.org/eng/assets/files/other/irrc_864_icrc_geneva.pdf

[10] Declaration Renouncing the Use, in Time of War, of Explosive Projectiles Under 400 Grammes Weight, St Petersburg, 29 Nov / 11 Dec 1868.

[11] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. "Intriguing properties of neural networks" arXiv:1312.6199, 2014.

[12] N. Akhtar, A. Mian, Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey, arXiv:1801.00553, 2018.

[13] I. Evtimov, K. Eykholt, E. Fernandes, T. Kohno, B. Li, A. Prakash, A. Rahmati, D. Song. "Robust Physical-World Attacks on Deep Learning Models" arXiv: 1707.08945, 2017.

[14] J. Lu, H. Sibai, E. Fabry, D. Forsyth "Standard detectors aren't (currently) fooled by physical adversarial stop signs". arXiv:1710.03337, 2017.