

“Ethical but Lethal” – The UK Position on AI Safety and Autonomous Weapons Systems

Naomi Colvin · Programme Director,
Blueprint for Free Speech, UK

28 November 2024



DISRUPTION
NETWORK
INSTITUTE

INVESTIGATING THE KILL CLOUD

Information Warfare, Autonomous
Weapons & AI

© The Author(s) 2024

Director & Editor: Tatiana Bazzichelli

Creative Producer & Layout: Jonas Frankki

Copy Editor & Proofreading: Janine Römer

Advisory Board:

Khalil Dewan, Anthony Downey, Thomas Drake, Jennifer
Gibson, Alexa Koenig, Matthias Monroy, Jesselyn Radack,
Andreas Schüller.

Disruption Network Institute

disruption.institute

Disruption Network Institute is a project
of Disruption Network Lab e. V.

disruptionlab.org

“Ethical but Lethal” – The UK Position on AI Safety and Autonomous Weapons Systems

Naomi Colvin · Programme Director, Blueprint for Free Speech, UK

Published 28 November 2024



This research focuses on the development of autonomous weapons systems (AWS), together with attempts to limit that development and use. It looks at the UK government’s position on AWS both domestically and internationally, particularly in relation to the country’s broader AI strategy.

Naomi Colvin is UK/Ireland/Belgium Program Director at Blueprint for Free Speech. She has a particular interest in whistleblowing as a freedom of expression issue and the intersection of digitally-mediated whistleblowing with the criminal law. Before coming to Blueprint, Naomi ran high-profile advocacy campaigns in this area. She occasionally writes for Byline Times. Naomi holds a Master’s Degree in European Political Economy and an undergraduate degree in Russian Studies, both from the London School of Economics.

Part One: AI safety in the UK

The 1st of November 2023 was an important date for UK policy and prestige in the area of Artificial Intelligence (AI).¹ From a six-month standing start, the British Government used its convening power to bring together 100 industry figures, academics and a small sprinkling of civil society representatives to Bletchley Park, the storied home of the Enigma codebreakers. The main topic of discussion was the potentially catastrophic consequences of unconstrained development of ever-more-powerful AI systems that could eventually evade human control entirely.

Two days of discussions and roundtables were held behind closed doors, with a set of public talks in London and Manchester to make up for the limited civil society participation in the summit itself. As if to provide some necessary spectacle the second day concluded with Tesla and X owner Elon Musk being subjected to a fawning interview by then-UK Prime Minister Rishi Sunak, live on YouTube (Sunak, 2023).

1 AI is the development of computer systems that can simulate cognitive tasks usually associated with humans. Contemporary discussions about AI largely reflect developments in machine learning using neural networks, in which systems trained on large amounts of data detect underlying patterns and use this to make predictions. Such systems are capable of improving their own performance without explicit human intervention.

In a year where large language models (LLMs) like ChatGPT² and the potential future applications of AI were a major focus of public attention, the Bletchley Park Summit was treated as a significant event by the UK media, marking a rare moment of British influence with Silicon Valley and a guest list to match.

The potential for AI to pose an existential risk to humanity had only recently become a touchstone of UK technology policy; it had been little-mentioned prior to May 2023. A White Paper – that is, an initial proposal for legislation – on AI published in February 2023 outlined the previous policy stance, setting the regulatory priority as one of encouraging innovation rather than averting extreme risks (DSIT and Office for AI, 2023).

The policy change, when it came, was abrupt. Investor Ian Hogarth wrote in the FT on 13 April 2023 that “we must slow down the race to godlike AI,” telling readers that AI engineers were now convinced that the advancement in frontier models³ was such that government intervention was necessary (Hogarth, 2023). The advent of artificial general intelligence (AGI), with competencies qualitatively superior to human beings, was on the way. Just two months later, Hogarth had been recruited to chair the UK’s new Frontier Model Taskforce, later to be renamed the AI Safety Institute⁴ (DSIT, 2023).

The pivot came as a surprise to many working in government, where policy on AI had not been treated as a major priority, still less one concerned with averting existential risk. Notwithstanding research work being conducted at British universities and a major US-owned AI lab – Google DeepMind – being based in London, the UK is not a leading player in the development of machine learning technologies globally.⁵

The Bletchley Summit, coming just six months after the adoption of the new policy agenda, was designed to establish the UK’s leading role in the rapidly developing niche of AI Safety. Investment of £100m was dedicated to the new AI Safety Institute, in order to develop techniques for the external supervision and audit of AI systems or models, previously regarded as a set of “open research problems.”

- 2 A LLM is an AI system trained on an enormous amount of textual information that predicts the next word in a sentence. These systems can produce surprisingly credible-sounding linguistic outputs and have particular strengths in areas like translation. Examples of LLMs include Open AI’s GPT-4 and Anthropic’s Claude.
- 3 A frontier model is one that pushes the boundaries of what is currently possible. In practice, the term has come to mean a model produced by one of the major US labs (Open AI, Anthropic, Google DeepMind, Facebook AI Research among others). Such models are often, but not necessarily, LLMs.
- 4 The organisation that became the AI Safety Institute was initially named the Foundation Model Taskforce, before being renamed the Frontier AI Taskforce and renamed once again in the run up to the Bletchley Park Summit in late 2023.
- 5 There are a number of different indices ranking countries in significance of their domestic AI industries. The UK typically ranks much higher in rankings that place greater weight on academic research than financial investment (Chakravorti, Bhalla and Chaturvedi 2023, Chopra, C., Kasare, A. and Gupta, P., 2024).

To some extent, the UK has been successful in this aim. The US set up its own AI Safety Institute in February 2024, housed within NIST, which cooperates with the UK on model evaluations and, unlike the UK, has concluded formal agreements with two of the major AI labs on sharing internal information (NIST, 2024). As of summer 2024 the UK Institute also has a liaison office in San Francisco (DSIT, 2024). Additionally, the May 2024 AI Summit in Seoul, intended as a follow up to Bletchley, concluded with an agreement to set up a network of AI Safety Institutes in jurisdictions including Canada, Singapore, the EU and Japan (Petrupoulos, 2024).

UK policy in this area has not been significantly impacted by the 2024 change in government. Despite private pre-election briefing that voluntary cooperation between the UK government and the major AI labs would be put on a statutory basis – something that formal cooperation with the US AI Safety Institute may have made unnecessary – Labour has continued with the Sunak government's reluctance to restrain the major AI labs through legislation.⁶

Frontier systems do seem to deliver new capabilities and those within the field generally agree that external scrutiny is necessary (Anderljung et al, 2023). On the whole, the dedication of public sector resources for this purpose, which brings a modest degree of transparency to the major AI labs, is a positive development.

Nevertheless, the direction the UK has taken is not ideologically neutral and this forms the point of departure for this essay. AI Safety encompasses a distinctive set of concerns, promoted by a specific group of people, which is closely associated with the major AI labs and Silicon Valley. That agenda is by no means representative of the entire range of concerns about the real-world impacts of AI, particularly when it comes to the ways in which the routine operations of systems affect the lives of individuals. This was a major source of criticism about the UK government's mid-2023 change of policy direction (McBride, 2023).

AI Safety as a body of ideas, as a doctrine, systematically excludes certain areas. This essay will argue that the UK's reluctance to apply its newly-found AI safety thinking to systems in the military domain is particularly striking, not least since this serves as a clear example of exactly how advanced systems can – and do – cause significant real-world harm and disruption on a geopolitical scale. Nor is this an area the UK public has no concerns about: polling carried out in 2023 shows that AI in weapons systems is ranked second only to autonomous vehicles as an AI-related source of worry for British citizens (Ada Lovelace Institute and the Alan Turing Institute, 2023).

Exclusions at the level of theory have followed through into policy. The AI Safety Institute, which sits within the Department for Science, Innovation and Technology (DSIT), has no remit to audit AI systems used by the UK military, and the Ministry of Defence (MoD) makes no commitment to subject its own AI systems to the rigorous

6 At the time of writing, media reports suggest that the current UK Government is planning to introduce a presumption of access to creative works as training data (Gross and Thomas, 2024).

and technologically innovative scrutiny the UK now considers important for civilian applications.

In fact, safety related initiatives in the military domain have taken an entirely different track, primarily focusing on achieving an international agreement on a subset of military technology, autonomous weapons systems (AWS). There is now over a decade of international arms control activity on AWS, in which the UK has played a significant – if not particularly constructive – role.

Neither the UK's domestic nor international policy on AWS has been subject to rigorous scrutiny, even in Parliamentary fora. One important exception was the House of Lords Special Inquiry on AI in Weapons Systems which held a series of hearings through 2023 and delivered its final report in December that year (AI Committee, 2023). While the conclusions of the committee did not achieve widespread media attention, the hearings were revealing, and the committee came up with some surprisingly sharp criticisms.

Nevertheless, all of these perspectives have distinct shortcomings when it comes to the reality of military AI and networked warfare as we see it in 2024. The final part of this essay looks at these tensions through a series of three case studies – humans in the loop in the civilian domain, safety recommendations in the military domain and finally Israel's use of AI targeting systems in Gaza, which brings the current predicament into sharp relief.

In the long term we are all dead

The UK's turn towards AI Safety as its keystone technology policy surprised many, both in and outside government. Some saw it as a form of state capture by Silicon Valley interests (Clarke, 2023). The change of focus was accompanied by unusually sharp public criticism of ongoing state-funded work on the real-world impacts of AI systems; the institutions affected by this were not ones on the fringes of official debate.

The Alan Turing Institute, set up in 2015 as the UK's national centre for data science and artificial intelligence, was heavily criticised by some close to government for being “blindsided” by the advent of LLMs. The Tony Blair Institute, in a report that was prominently featured in the national press, went so far as to argue that the Turing Institute had “failed” and should be “defunded entirely” (Sellman, 2023; Tony Blair Institute, 2023).

At the same time, other existing institutions in government, such as the AI Council, made up of some of the UK's leading academic experts, were abruptly disbanded. Professor Wendy Hall, who had been on the disbanded Council told the House of Lords Communications and Digital Committee in October 2023: “To be honest, I could not tell you whether there is a functioning AI strategy” (Communications and Digital Committee, 2023a).

AI Safety is more than a specific area of concern: it is a discernible assemblage of ideas about the existential risks posed by AI that emerges in Silicon Valley and has been disseminated through Effective Altruist, Longtermist and adjacent networks. The prospect of AI systems that have superseded human capabilities and have the capacity to evade human control – or, on a less dramatic level, the difficulty of aligning systems with less extraordinary capabilities with human interests – is seen by this group to be one of the most serious problems facing humanity.

A significant amount of effort that has gone into movement building around these concerns, particularly at the UK's most prestigious universities. Regular free courses and reading groups are held online and in-person, all aimed at bringing people into the sector and finding them employment. The syllabuses produced for these courses are a good representation of what the agenda-setting part of this community at least perceives as important (BlueDot Impact, 2024).

In AI Safety terms, the emphasis placed on human control of AI means that development at the frontiers is invariably framed as a national security issue, with policy proposals situated in the realm of foreign policy and grand multilateral initiatives. In effect, the approach is one of state supervision of AI development, overseeing and sharing the risk borne by AI developers, who are themselves grappling with the hard technological problem of adding safety guardrails to their systems.

Post-1945 efforts to control the proliferation of nuclear weapons – particularly multilateral treaties with an enforcement mechanism – are frequently cited by this group, not just as an analogy, but an outright recommendation. This is despite the overall international context, given that we have seen the deterioration not only of nuclear non-proliferation arrangements in recent times, but also the complete collapse of the international arms control enforcement regime in chemical weapons.

In the UK, this focus on AI as a national security issue has found a receptive audience in government. As Ian Hogarth told the House of Lords Digital and Communications Committee in September 2023:

“There are experts in the British state who have devoted their lives to protecting us from cyber and biological attacks. We are trying to synthesise the expertise we are bringing to the task force in frontier AI research with the existing expertise inside the British state on national security frontier challenges. In some ways, that is expressed by the work that we are doing on the ground as well as on our advisory board” (Communications and Digital Committee, 2023).

Among the members of the External Advisory Board of the AI Safety Institute are the Director of GCHQ and the UK's Deputy National Security Advisor and Chief Scientific Advisor for National Security, both roles based in the Cabinet Office (DSIT, 2023a). This information is not available on the AI Safety Institute's public-facing website, which was published in May 2024, which simply states that “our advisory board

comprises national security and machine learning leaders” (AI Safety Institute, 2024).

Despite the frequent appearance of national security and arms control agreements in the AI safety literature, concern about the practical application of AI technology in the military among prominent writers is low, with a very few exceptions.⁷ Some key thinkers – for instance, Nick Bostrom, whose academic thought experiments have been extremely influential in this community – explicitly state that military AI is part of a different (implicitly less important) discussion, effectively downplaying it as an area of concern: “these are not unreasonable things for some people to be thinking about” (Bostrom, 2014, p. 323).

In general, comments on the military practices of state parties have been restricted to analysis of ongoing discussions about limiting the development and proliferation of AWS, where the primary driver has been civil society and the arms control community. An open letter from 2016 registering opposition to the development of AWS from robotics and AI researchers was unusual in this regard. Interestingly, the letter, organised by the Future of Life Institute, which Nick Bostrom founded, did not include his own signature (Future of Life Institute, 2016).

Where military concerns do appear in the AI safety literature, these are largely looked at a foreign policy perspective, in terms of how differential patterns of AI advancement and first mover advantages could destabilise the international order. The need for government to step in and “synchronise” the speed of development between private companies or even countries has been a feature of these discussions (Bostrom, 2014, p. 161). There may be a connection to be drawn with this pro-intervention argument and the Spring 2023 calls from the AI labs for governments to take potentially dangerous progress in AI more seriously.

There is also likely a link between AI Safety’s relative lack of interest in real-world applications of AI in the military and its more general elision of concerns about the concrete functioning of real-world systems. Save for national security-adjacent issues such as the potential for LLMs or systems like DeepMind’s AlphaFold to enable non state-actors to produce nuclear or biological weapons, it is hard to locate much thinking on the ways the operation of systems impacts those at the receiving end.

In general, writing on AI Safety – while valuable in many ways – shows a striking lack of curiosity about structural injustice, the security state and even corporate governance. The assumption that commercial entities are good actors, is probably no longer sustainable, even within this community.

In particular, there is a growing recognition that whistleblowing at the major AI labs is likely to prove important, as the recent Right to Warn open letter, written by a group of former Open AI employees in the wake of a series of departures and

7 Both Stuart Russell and Paul Scharre have engaged seriously with international standards-setting around autonomous weapons systems, though they have contrasting attitudes about the possibility of an international arms control treaty.

management crises demonstrates (Right to Warn, 2024; Kototajlo and Ball, 2024). Among the positive elements of California's comprehensive AI regulation Bill SB 1047 – which was eventually vetoed by Governor Gavin Newsom in October 2024, reportedly in response to lobbying from the major labs and investors – was specific provision for AI whistleblowing, including recognition of the necessity of anonymous disclosure for those in the sector.

There are some obvious structural explanations lying behind the biases in what AI Safety considers most important, not least the relationship of major US technology corporations with the national security state. But there are also tendencies we can identify at the ideological level. Effective altruism is a call for those who have benefited from the existing economic model, primarily but not exclusively in technology, to tithe part of their income to causes selected on the utilitarian basis of what will benefit the greatest number of individuals. This means, for example, that global health or vaccination programmes, which work at the population level and can be assessed quantitatively, tend to be a favoured cause in this sector.

Longtermism is a mode of thinking closely linked with effective altruism, often associated with Oxford philosophers Toby Ord and William MacAskill. Those who subscribe to this programme generally discount sources current of hardship in favour of tackling one of a few issues that have been identified as major existential risks for humanity in general. These issues typically include the prospect of uncontrollable AI as well as pandemics (Crary, 2023).

Both Effective Altruism and Longtermism as approaches therefore tend to be sceptical of the value of action to remedy structural injustices in the present, or that which adopts a critical approach to current structures of power generally. In the world of animal welfare, where Effective Altruism's interest in promoting veganism has led to associated organisations becoming the dominant funders in the sector, this has caused significant problems for established organisations or activists, who seek to remedy current animal suffering or focus on systematically disadvantaged populations (deCoriolis et al., 2023; Halteman, 2023).

It is not coincidental that the people most likely to face the consequences of the inequitable, inflexible or otherwise unjust operation of algorithmic systems are also those disadvantaged by existing power structures; among them benefit claimants, migrants and – in the context of military AI – civilians living in combat zones. Some observers have seen an embedded racism in Longtermism as a theory, as well as outright prejudice in some of its individual proponents (Crary, 2023; Gebru and Torres, 2024).

British policy on AWS

Attempts to establish rules in use of AI in the military long predates the advent of LLMs and a primary area of concern has been the integration of autonomous

functionality in weapons systems. Since 2012, the development and regulation of autonomous weapons systems, or lethal autonomous weapons systems (LAWS), has been the subject of concerted discussion in international arms control fora. Given the exclusion of defence matters from other multilateral agreements on AI, for instance the EU AI Act and the Council of Europe Framework Convention, negotiations around AWS still look like the likeliest route for binding legal rules in this area.

International discussions on AWS were initially conducted through the Certain Conventional Weapons (CCW) forum at the UN in Geneva, which requires state parties to agree on a position unanimously, and have more recently progressed under the aegis of the UN General Assembly. In November 2023 the UN General Assembly adopted Resolution 78/241 on 'lethal autonomous weapons systems', with an enormous majority of 152 states voting in favour of the resolution. Campaigners on this issue assess that a majority of the world's states are now in favour of an international treaty, discussions about which are due to continue during the 2024 UN General Assembly session (Automated Decision Research, 2024).

As is often the case in international negotiations, multiple definitions of AWS have been proposed over the course of the past decade. These definitions have often been put forward by state parties with an interest in limiting the scope of arms control measures. The term LAWS is particularly vulnerable to the use of maximalist definitions that would exclude many types of weaponry that pose a challenge to humanitarian norms.⁸

Nevertheless, at this point in time, for the purposes of proposed controls there is a reasonable amount of consensus around the general definition used by the International Committee of the Red Cross (ICRC), which was also employed in a rare joint declaration by the UN Secretary General and the ICRC President urging progress towards an international agreement by 2026:

"Autonomous weapon systems select and apply force to targets without human intervention. After initial activation or launch by a person, an autonomous weapon system self-initiates or triggers a strike in response to information from the environment received through sensors and on the basis of a generalized "target profile". This means that the user does not choose, or even know, the specific target(s) and the precise timing and/or location of the resulting application(s) of force" (ICRC, 2021; ICRC, 2023).

The broad thrust of proposals for a treaty is that the use of AWS cannot be brought into compliance with principles of international humanitarian law (IHL) in the absence of new legal rules. Among those principles are the need to distinguish between combatants and civilians, the proportionality of any military action, its necessity in military terms and the need to take reasonable precautions to minimise harm and prevent errors. IHL dictates that state parties and their militaries need to be accountable for abiding by these principles (ICRC, 2024).

⁸ As such, AWS is my preferred term for the purposes of this essay.

The nature of AI systems, which produce probabilistic (and potentially non-replicable) responses, creates problems of predictability, which renders questionable whether they can be subjected to meaningful human control. While there are potentially technological safeguards that can be applied to ameliorate this issue, advocates of new legal rules around AWS recommend that unpredictable systems, as well as those which target human beings, should be banned outright. Systems that lie outside these categories should be subject to defined limits on what they target, how they are deployed and the degree to which they are subject to human supervision.

The definition of AWS used by treaty proponents has its limitations: the use of AI to determine the targets of missile strikes, such as has been seen in current Israeli operations in Gaza, likely falls outside the definition completely. This example is the subject of the third case study in this essay.

Multiple examples of AWS already exist and have been deployed in theatres of conflict. Indicative lists of existing systems have been produced by the House of Lords Special Inquiry and Stop Killer Robots, part of the civil society coalition pressing for an arms control treaty (AI Committee, 2023, p. 93; Automated Decision Research, 2024a). These lists are not fully comprehensive – they do not, for example, take note of the rapid evolution of drone technology in Ukraine, where US investors including former Google CEO Eric Schmidt have been funding the development of low-cost drones with the capacity to fly autonomously, countering the impact of signal jamming (Hudson and Khudov, 2023). Schmidt's own company White Stork is now developing autonomous drone technology in the US, primarily for use in Ukraine (Emerson and Nieva, 2024).

The UK has not been a proponent of a new arms control treaty and has identifiable interests in not being so. Not only does the UK have a significant arms industry (the Stockholm International Peace Research Institute, or SIPRI, ranks the UK as the world's 7th biggest arms exporter), the country is one of the key developers and sponsors of AWS, particularly autonomous drones. As with machine learning generally, academia provides a major contribution to UK capabilities.

A 2022 survey of AI and autonomy research looked at research agendas at 13 UK universities that had accepted funds either from the UK MoD or from arms manufacturers. The researchers identified 65 separate research programmes with clear potential application to AWS, with 17 of these programmes appearing to be directly related. The apparent dependency of universities on these funds means that the ability of individual academics to act according to their conscience is limited (UK Campaign to Stop Killer Robots, 2022).

Several recent examples of British AWS development have become visible through the AUKUS defence pact between the US, the UK and Australia, of which sharing of military AI technology is a part. Publicised trials include a trial of a "collaborative" drone swarm in April 2023 with the claimed ability to identify targets, which could be used by all three allies (Skelton, 2023; DSTL, 2024).

The UK is one of ten countries that researchers consider to be opposed to an international treaty to limit AWS, alongside AUKUS partners the US and Australia, as well as Israel, India, Japan, South Korea, Estonia, Poland and Russia (Automated Decision Research, 2024a). Despite the CCW process having exhausted itself in the eyes of observers, the UK's current position, as explained by then Procurement Minister James Cartlidge MP to the House of Lords Special Inquiry, is that the CCW remains the "most appropriate" forum for discussions about AWS, within the bounds of current IHL, as opposed to the UN General Assembly – which is more likely to want to settle new legal rules (AI Committee, 2023). This a position shared by the US and Australia as well as India and Russia.

The UK's very definite stance on arms control in this area has not tended to be explicitly justified to Parliamentarians, let alone in other public fora. Indeed, the UK government's official statements on this issue – as well as communications with campaigners – generally seem designed to create a sense of ambiguity. Government ministers refused to confirm their position on UN General Assembly Resolution 78/241 until days before the vote.⁹

Voting on Resolution 78/241 coincided with the Bletchley Park Summit. Media briefings promised pre-Summit ministerial announcements on the UK's position on military AI, which did not in fact come to pass. However, while military domain AI was not on the formal agenda, it was not entirely absent from the Summit. US Vice President Kamala Harris used a press call to highlight the US-led Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy, which the UK had endorsed, again without a public announcement (US Department of State, 2023).

Originally unveiled in February 2023, the Political Declaration is a much weaker instrument than any proposed treaty and is framed in largely advisory terms. Notably, it does not place limits on the development of – or even define – AWS, but simply urges states to "take appropriate steps, such as legal reviews, to ensure that their military AI capabilities will be used consistent with their respective obligations under international law."

States are enjoined to take measures to prevent bias and errors, embed their systems within "a responsible human chain of command and control", and provide training so that those charged with operating systems can "make appropriate context-informed judgments." Having a sufficiently-trained human in the loop in this view, then, represents a sufficient safeguard that brings automated weapons systems into compliance with existing IHL.

Domestically, policy documents published by the UK MoD, agree that no new binding rules are necessary: "nothing about AI fundamentally changes our obligations under UK law and international law, or the importance we attach to the

9 <https://ukstopkillerrobots.org.uk/wp-content/uploads/2023/10/UNGA-Resolution-on-Lethal-Autonomous-Weapons-Systems-UK-Government-responds.pdf>

standards, values and norms of the society we serve.” A human in the loop represents the most important safeguard: “we are very clear that there must be context-appropriate human involvement in weapons which identify, select and attack targets” (MoD, 2022, p. 1; MoD, 2022a).

The discrepancy between the UK’s position on civilian and military AI was noted by the year-long House of Lords Special Committee of Inquiry on the Use of AI in Weapons Systems, which examined the UK’s position on AI in the military domain, interviewing 35 witnesses over 15 separate evidence sessions. The Committee’s final report, *Proceed with Caution*, was published in December 2023, a month after the Bletchley Summit, and was highly critical of elements of the UK’s position (AI Committee, 2023).

The Committee members were particularly critical of the MoD’s failure to agree on even a provisional definition of AWS. In debates following the publication of the report, committee members were explicit that they saw this as obstructive both in terms of potential international agreements and for basic domestic accountability:

“How can we actively seek to engage with policy in regulating AWS if we cannot find even provisional words with which to define it? It is like attempting to make a suit for a man whose measurements are shrouded in secrecy and whose very existence is merely a rumour. These are, of course, enormously complex questions but in making good policy, complexity should not be a refuge but a rebuke. It is the job of Governments of any political stripe to be able to articulate their approach and have it tested by experts and dissenting voices” (HL deb, 19 April 2024).

The Committee also called for the UK government to explicitly support international efforts to limit AWS, that they “should step up and show leadership” (AI Committee, 2023, para. 203), and apply the standards articulated in the Declaration that closed the Bletchley Summit to be applied to the military domain.

Moreover, the Committee called for an outright ban on the use of AI in nuclear communication, command and control systems, and a maintenance of human control, which is “essential ... for moral agency and legal compliance” though seemingly rather hard to define in practical terms (AI Committee, 2023, para. 175).

In minutes that were published as a result of the Committee’s work, it appears that the MoD’s own AI Ethics Committee concurred with the authors of the report that what represents ‘meaningful human control’ is extremely challenging to define in any given context.

“Further work is required to understand what context-appropriate meaningful human control looks like for different AI systems with different degrees of autonomy” (MoD AI Ethics Advisory Panel, 2024).

Notwithstanding its potentially newsworthy criticism of government policy, the Special Inquiry report was published to little media attention. In the short term,

perhaps the most concrete impact of the report is that UK government definitively ruled out supporting a treaty, in Parliament:

“It is foreseeable that, in international negotiations, those who wilfully disregard international law and norms could use a definition to constrain the capabilities and legitimate research of responsible nations. It is also for that reason that, after sincere and deep consideration, we do not support the committee’s call for a swift agreement of an effective international instrument on lethal autonomous weapons systems — that would be a gift to our adversaries” (HL deb, 19 April 2024a).

There does not appear to have been any change in the UK position on AWS under the Labour government that took office in July 2024.

Part Two: Case Studies

1 – Humans in the civilian loop

For clarification on what “meaningful human control” might look like in practice, it is worth looking at studies of similar devices in civilian AI. The human in the loop is often proposed as a panacea to potential inequities and biases in the operation of algorithmic systems. Algorithmic systems do not necessarily use machine learning technologies to produce their results – that is to say, they are much more likely to produce consistent results from a given set of inputs and do not ‘learn’ in the same way that AI systems do. Nevertheless, from the perspective of the end user, or the individual obliged to interact with such a system, the experience of being subject to a machine-produced decision rather than the contextual consideration of a human-decision maker is broadly similar.

There are still some exacerbating factors that come into play with machine learning. AI systems, and LLMs in particular, by their very nature tend to reproduce patterns and biases present in their training data. As such, existing injustices are likely to be perpetuated in future decision-making. More than that, the inherent unpredictability of outputs from machine learning systems means that the process by which outcomes are arrived at is hard, if not impossible, to understand in retrospect.

In some areas, as a remedy to these sorts of problems, the presence of a human in the loop is mandated by law. Article 22 of Europe’s 2018 General Data Protection Regulation (GDPR), for instance, gives data subjects – those whose information is processed in the operation of system – to “obtain human intervention on the part of the controller” as part of their rights of redress.

Whether Article 22 has proven effective in practice, either as a means of protecting rights or as a remedy for breaches of rights, is debatable. For one thing, many of the most notorious cases internationally of algorithmic decision-making

leading to injustices – for instance the racial biases in a US criminal justice risk assessment system described by *ProPublica* in 2016 – were not situations in which there was no human involvement. Even when systems are only intended to provide “decision support”, the extent of human intervention or supervision of those decisions can be extremely minimal. The problems are likely to be magnified in cases where the decision at issue is complex or nuanced (Edwards and Veale, 2017, p. 45, 51).

Consideration of the role of ‘humans in the loop’ has usually come from researchers with an interest in equity or fairness concerns in the operation of current, as opposed to future, speculative, systems. Within this group, there is a divide between those who consider that some amelioration of problems is possible using technical means – for instance, with the careful curation of training data to eliminate pre-existing bias, or the introduction of metrics to identify likely types of unfairness such as discrimination by gender, age or location.

The more radical account of these issues sees AI systems as embedding structural systems of oppression, potentially even by design. In this account, bias in training data is not only a risk for perpetuating structural injustice, “it obscures the possibility of a structural break with the past.” Embedding these values in technology gives them a sense of timeless authority and consistency, “the aura of science.” In this sense, AI systems are profoundly undemocratic as they make it more difficult to challenge power and enact change in the real world (McQuillan, 2022, p. 43, 51).

Certainly, the employment conditions of those engaged to be humans in the loop, content moderators and similar functions are infinitely poorer than those of the developers whose work they are ostensibly supervising.¹⁰ Their individual agency, too, is denigrated, treated as “interchangeable sets of eyeballs” (McQuillan 2022, p. 14).

Perhaps surprisingly, the radical critique of AI has something in common with the most technocratic readings of AI Safety, certainly insofar as the role of a human in the loop is concerned. Both Dan McQuillan and Nick Bostrom identify dynamics whereby a human in the loop will predictably be rendered less and less effective until, eventually, they are dispensed with entirely. For McQuillan, this is foreseeable precisely to prevent the human in the loop playing the role they are intended to:

“With humans in the loop, the smooth functioning of institutional processes that are discriminatory, unethical or unjust are always vulnerable to acts of individual conscience or collective refusal. With AI alone no risk exists” (McQuillan, 2022, p. 38).

He suggests that where scandals emerge in the operation of AI systems, particularly those deployed in the public sector, the human being involved in implementation will inevitably take the blame. The almost inevitable consequence of

¹⁰ In recent years, Foxglove and other organisations have done much to highlight the deeply inequitable working conditions experienced by moderators for Meta and the other internet giants, many of whom are located in the global south (Rowe, 2023).

this is that a few scandals down the line, human involvement will be dispensed with in the interests of safety.

It is striking that the likely trajectory for humans in the loop laid out by Nick Bostrom in *Superintelligence* is actually very similar. He suggests that errors made by AI systems that are “precursors” to AGI will likely lead to demands for ‘safer’ systems; that is, ones that exclude the fallible human element. He employs the example of a drone attack on civilians. Combined with the predictable race to develop ever more powerful systems, he doubts whether the continued presence of a human in the loop is feasible or credible (Bostrom, 2014, p. 143).

The identification of similar types of dynamics does not lie far under the surface of discussions about AI safety in the military domain, which primarily rely on the elusive concept of “meaningful human control” over AI-enabled systems.

2 – The military take on AI Safety

While largely unexplored within the AI safety world, there has been no shortage of discussion about how AI could change warfare and the ways militaries operate. Part of this discourse is practically analogous to the catastrophic risk discussions so prevalent within AI safety, with speculation about the potential ramifications of AGI on the level of grand strategy. Nevertheless, these discussions are not solely of the longtermist type, and speculations about AGI have been accompanied by more concrete, and immediate, concerns.

While AWS do feature in these debates from time to time, this has been more of a primary focus for civil society and the international humanitarian community. In fact, long-form analyses are surprisingly thin on the ground, with Paul Scharre's *Army of None* forming something of an exception. Among military theorists, discussion tends to be broader than battlefield technology, taking in the impact that AI is likely to have on strategic, logistical and intelligence functions.

Writing on the military dimensions of AI tends to have a dual focus: firstly on the potential for AI to give military forces an advantage over their opponents, and – as a secondary consideration – the implications for accountability and IHL. Significantly, several observers in this area have acknowledged that these two dynamics are not easily reconcilable with each other.

Theorists in this area tend to disagree on how significant military domain AI is likely to be. The maximalist argument in this realm is exemplified by the concept of hyperwar. The term was coined in 2017 by General John R. Allen, U.S. Marine Corps (Retired), and Amir Husain:

"In military terms, hyperwar may be redefined as a type of conflict where human decision making is almost entirely absent from the observe-orient-decide-act (OODA) loop. As a consequence, the time associated with an OODA cycle will

be reduced to near-instantaneous responses. The implications of these developments are many and game changing" (Allen and Husain, 2017).

The authors speculate about a variety of devices that would affect hyperwar in the physical world, from autonomous single devices to swarms, but all controlled by a 'synthetic intelligence' analogous to AGI that is superior to human cognition. As with discussions of AGI, there has also been discussion about the implications of the perception of actual or incipient technological superiority for the prevailing balance of power (Goldfarb and Lindsay, 2022).

The concept of hyperwar is essentially Bostrom's Superintelligence applied in the military context. As with civilian applications, the three dimensions identified by Bostrom in which AGI could supersede human capabilities – (speed, cooperation and the advent of abilities humans simply do not have¹¹ – are also relevant to lessgod-like applications of AI in the military domain.

In a military context, coordination could apply to drone swarms or the less visible consolidation of information across military systems to provide something like strategic-level assistance. Quality relevant to military applications could mean, for example, deriving entirely new types of insight from surveillance data.

Nevertheless, speed is the dimension in which the opportunity for military advantage is most obvious: reducing the length of the "eye to fire cycle" clearly offers an advantage over an adversary, just so long as there is a gap in technological capabilities between the two. Similar dynamics could just as clearly apply in other military functions, such as the processing of intelligence.

Paul Scharre has noted that speed on its own could potentially lead to disaster if a human is not involved. In *Army of None*, he uses the example of Stanislav Petrov, whose role "in the loop" of a nuclear warning system was arguably one of the most critical moments of the Cold War. While the time involved in actually moving physical materiel leads Scharre to downplay the possibility of unintentional outbreaks of conflict in the physical realm, he accepts that this is a real and present threat in the context of cyberwar. In addition:

"Even though accidents with physical autonomous weapons will not cascade into all-out war in mere seconds, machines could quickly cause damage that might have irreversible consequences" (Scharre, 2019, p. 220).

The fog of war

It is not the case that there is a consensus about the military potential of AI. For instance, some see AI primarily as a general purpose advance of primarily economic

11 Bostrom refers to this third category as "quality," which is potentially open to confusion. He uses this in the sense is of new qualities of cognition primarily rather than higher quality judgement in areas where humans are already competent - though of course the two are related (Bostrom, 2014, p. 68-70).

importance (Ding and Dafoe, 2023). Even among proponents, however, there are other grounds for scepticism about the adoption of military AI.

Prime among these is AI's dependence on large datasets, both for training and operations. In large part for this reason the UN Institute for Disarmament Research, based on a series of interviews and workshops with state parties, concluded that the primary military application of AI over the next ten years was likely to be in upstream processes, such as the processing of intelligence data rather than weaponry (UNIDIR, 2023).

It has been argued, with some justification, that those writing in this area have tended to downplay the potentially devastating impacts of AI-enabled intelligence, surveillance, target acquisition, and reconnaissance functions (ISTAR), which have all tended to be regarded as less problematic than AWS (Khlaaf, West and Whitaker, 2024). The role that AI-supported targeting has played in Israel's 2023-24 operations in Gaza brings this into sharp relief.

Nevertheless, in the 'insufficient data' analysis, battlefield conditions are seen as unlikely to produce data of adequate quality, both because they are inherently unstructured and hard to predict, not least in the absence of conditions of intense pre-existing surveillance (Goldfarb and Lindsay, 2022). There an analogy to be drawn here with civilian self-driving vehicles, which have to this date proven unable to operate reliably or safely in much more predictable urban environments.

The acquisition of data itself, alongside attempts to degrade data available to opponents, is therefore critical and likely to become an area of competition in its own right. The crowdsourcing of Russian troop positions in Ukraine is arguably a current example of this (Feldstein, 2024). Indeed, the UK Army refers to data as a "critical asset" (MoD, 2023).

A related concern is the lack of good data due to the rarity of the events concerned. Armed conflict can be a relatively infrequent event for most militaries and data in related realms – for example, nuclear decision-making – is even harder to come by. The technological 'fix' for this issue is the reliance on synthetic, ie. machine produced, data, which inevitably exaggerates pre-existing biases (Saltini, 2023).

A second area which gives rise to doubts about the practical use of AI in the military context is predictability, which is closely tied to issues of accountability as well as effectiveness. The probabilistic nature of machine learning systems, combined with the noted tendency of such systems to "hallucinate", or come up with outputs for which no data exists, means that a system may very well not do what an operator expects it to. How, then, can you attribute responsibility?

An inability to predict an outcome means that the ways in which systems could fail are also highly unpredictable, with potentially devastating effects, similar to the loss of control arguments prevalent in AI Safety discussions. There is some experimental evidence that AI systems exhibit greater volatility in their conclusions than humans faced with the same scenario, though this research was restricted to LLMs (Lamparth et al., 2024).

Unsurprisingly, the lack of certainty over the role AI could play has an impact on trust on the part of military personnel who could be asked to use or rely on it in future. A survey conducted by researcher Paul Lushenko among US officers attending the Army War College in Carlisle, Pennsylvania, and the Naval War College in Newport, Rhode Island, demonstrated stronger support for forms of AI that preserve a human role for ultimately authorising or effecting a decision (Lushenko, 2023).

Such results should nevertheless be seen through the lens of human tendencies to rely, to an ever-greater extent, on the outcomes of automatic decision-making through wider organisational imperatives, something that Lushenko himself signals.

"Despite supporting the adoption of such technologies to optimize decision-making at various levels and degrees of oversight, officers do not trust the resulting types of potential warfare on account of emerging AI-enabled capabilities. This result suggests that US officers may feel obliged to embrace projected forms of warfare that go against their own preferences and attitudes..."

Lushenko goes on to argue that adoption of new technology could be facilitated by connecting it with humanitarian functions to make it more acceptable. Others frame adoption as an absolute imperative in the light of geopolitical competition between the US and China or Russia's invasion of Ukraine (Scharre, 2023).

The suggestion that organisational imperatives, or indeed a moral belief in the need to apply this kind of technology in a given circumstance, can overcome inherent trust barriers or practical concerns is borne out to some extent in the context of Israel's current operations in Gaza, which are discussed later in this essay.

Reservations about applications of military technology are not restricted to military establishments, particularly given the dominance of the civilian sector in AI development. Surveys show that civilian workers in AI – being well aware of its biases and limitations – are notably more sceptical about the use of the technology for military purposes than the population on average. Many, of course, also have moral objections about associations with militaries, repressive governments or immigration controls. The ability of employees in the US to register objections, albeit at the risk of losing their jobs, has led to the cancellation of some projects such as Google's Project Maven and Dragonfly and has been identified as a factor with geopolitical consequences, particularly vis-à-vis China (Scharre, 2023, p. 22-23, 65).

Finally, even though the presence of a human in the loop, or "meaningful human control", is frequently deployed as a panacea for the IHL compatibility of military AI capabilities, there is a notable degree of scepticism about how this could work in practice. Notably, the presence of a human in the loop represents a limitation on the speed with which AI systems can operate – that is, it compromises one of military AI's primary selling points.

Some, such as the UN Disarmament Agency, have portrayed this tension as a potential positive in terms of compliance with IHL:

"Unless humans are completely removed from all these processes, ultimately the rapidity of the operational tempo will still be determined by humans and their roles in the observe, orient, decide, act (OODA) loop. In other words, we still have the possibility of controlling the tempo, as well as the use of AI" (UNIDIR, 2023).

However, others have identified a fundamental incompatibility and a likely dynamic that tends towards less, and less meaningful, human input over time. In fact, some commentators see a very limited role for human oversight after the design and building stage of military AI systems:

"Though it yields some benefit, the prescription of human intervention is too weak and too frequently counterproductive to consistently control automated systems where speed is at a premium.... The important lesson ... is that efforts to control these systems are vastly more effective at the time of design than they can be in decision making during operations. Early is imperative. Late is too late" (Danzig, 2018).

Even if the long-term involvement of humans in operational decisions is viable, as we have seen, what "meaningful human control" – or even "context-specific human control" – means in practice has proved very difficult to define, with both the House of Lords Special Inquiry and indeed the UK MoD's own expert panel admitting difficulties.

Attempts to square this circle raise more questions than they answer. Paul Scharre distinguishes two functions humans could be playing in the operation of military AI capabilities. Firstly, they could act as a failsafe that provides a sense check and performs a supervisory function, though the functionality of the system itself could happen without their input (Scharre, 2019, pp. 322-323).

Secondly, a human could act as "moral agent," making a value judgement about whether the use of force is appropriate – which more closely resembles the role a human operator is supposed to be performing to ensure accountability and compliance with IHL. These functions may overlap, in Scharre's opinion, but they do not necessarily have to – something which has been largely borne out by the experience of Israel's use of AI in its ongoing military operations in Gaza.

3 – Automated Targeting in Gaza

In November 2023 and April 2024, investigative reporting by Israeli news outlets +972 and *Local Call*, alongside the UK's *The Guardian* revealed the extent to which

AI was playing a role in Israel's military operations in the Gaza strip, initiated in response to Hamas' incursion into southern Israel on 7 October 2023.

These reports - which, importantly, included testimony from individuals tasked with interacting with these systems in the field - show how challenging the reality of AI-enabled warfare is, not least in its implications for existing systems of arms control and the likelihood of new limitations being instituted. These revelations also highlight the degree to which critical threats to human welfare elude the existing debates on both AI safety and AWS (Abraham, 2023; Abraham, 2024).

The use of AI by the Israel Defence Forces (IDF) had been reported previously, albeit not in significant detail. Indeed, the existence of several of the systems described by +972 and *Local Call* had previously been revealed by the IDF itself in limited public statements (Dumbi, 2023). Col Yoav, Commander of the IDF Artificial Intelligence centre, discussed new military capabilities at a January 2023 conference in Tel Aviv University, which was reported in local media.

"One of the significant tools we have built and run today is a system that knows how to find 'dangerous' people based on the importance of a list of people who have been dropped and entered into the system. This process is done in seconds, which in the past would have taken hundreds of researchers several weeks" (Ahronheim, 2021).

Lavender, the system described by Yoav, was developed within the IDF and first used to identify human targets in Gaza over the course of Israel's eleven-day Operation Guardian of the Walls in May 2021. Israel itself has dubbed Guardian of the Walls "the first AI war."

Recent reporting supplies much more information about how Lavender and other systems actually operate in practice. Lavender is a tool for developing targets that assesses the likelihood of whether a given individual is likely to be a member of Hamas or Islamic Jihad, of any rank. The system is trained on data about known operatives and uses this to rank unrelated individuals with a score between 1 and 100.

Human Rights Watch have speculated that Lavender involves "positive unlabelled learning" - that is, it is trained on a combination of tagged (ie. positively identified) and untagged data (Human Rights Watch, 2024). If this is indeed the case, the way the training data has been tagged is questionable and provides an extreme case of how pre-existing biases can be perpetuated by automated systems. One source told +972 that the training set included individuals like civil defence workers, who could not in any way be considered legitimate military targets.

While the purpose of Lavender is to mark out individuals, the related system Gospel (Habsora) identifies buildings and other physical locations for targeting, including private residences. Another system, Where's Daddy, correlates phone and location data to produce an alert when an individual is detected in their home. The different systems are not employed in isolation. Individuals identified by Lavender would be inputted, at scale, into Where's Daddy - effectively putting them under close

surveillance and marking their residence out for bombing when their presence is detected.

A member of the IDF Targets Administrative Division, which is tasked with developing and operating these AI-enabled targeting tools, told journalists that speed and the quantity of targets delivered were the main metrics they were judged on:

“We prepare the targets automatically and work according to a checklist... It really is like a factory. We work quickly and there is no time to delve deep into the target. The view is that we are judged according to how many targets we manage to generate.”

An absence of checks built into the system was accompanied by a significant loosening of safeguards at the level of policy. Sources told +972 and *Local Call* that, since October 7, residential buildings in Gaza were being bombed, often overnight, regardless of whether civilians were – or were likely to be – present, with no real attempts to warn or evacuate residents prior to a strike. The unguided missiles used in these strikes were known to be more likely to cause significant collateral damage.

Lavender and Gospel are not the only AI tools that have been developed. Other known systems include Alchemist, which provides real-time alerts to military personnel on the ground about potential threats, and Depth of Wisdom, which maps Gaza's tunnel network. Another system tracks population movements within Gaza and locations where people are gathering through analysis of mobile phone location data. In October 2023, the tool was used to track the evacuation of the population of north Gaza in advance of the IDF's advance into the territory (Human Rights Watch, 2024).

According to reports an additional and recently developed set of tools is available to IDF personnel in the field, including a YouTube-like interface ‘Z-tube’ for accessing live video footage, including that taken by drones (Abraham, 2024a). Other tools include a collaborative platform for marking the geographical location of targets in real time (‘MapIt’) and for detecting targets using AI to analyse patterns of behaviour (‘Hunter’).

While the precise data points used in systems like Lavender are impossible to define with certainty in the absence of access to source code, all these applications are built on top of existing systems of mass surveillance of Gaza's 2.3 million residents, which includes the routine acquisition of phone, satellite and biometric data.

The extent to which Israel has exploited detailed and individualised knowledge of Palestinian populations is also demonstrated by the Blue and Red Wolf facial recognition systems that operate at checkpoints in the Occupied West Bank territories (Dwoskin, 2021; Amnesty International, 2023).

If anything, the surveillance of Gaza's population has been even more extensive. While phone networks in the Occupied Palestinian Territories operate on 3G, which are vulnerable to interception by state actors, networks in the Gaza strip have been forced to maintain 2G connections, long superseded in most of the world, which

facilitate the mass collection of unencrypted audio in real time (Al Shabaka, 2024; Timberg and Soltani, 2013). In the opinion of Human Rights Watch, among others, the pervasive, “systematic” surveillance of the population of Gaza is itself a human rights violation (Human Rights Watch, 2024).

Data used in these systems is also drawn from the surveillance capabilities of Israel's allies, particularly the United States and the United Kingdom, ostensibly for the purpose of recovering Israeli hostages (many of whom were dual citizens) who were abducted on 7 October 2023. The US has been flying surveillance drones over Gaza from the earliest months of the conflict, and instructions issued to US intelligence officers stationed in Israel suggest that satellite imagery has been used to assist in targeting (Klippenstein and Petti, 2024).

RAF Akrotiri, the British Air Force base in Cyprus, has been used as a staging post for US Special Forces flights and weapons delivery (Kennard, 2024a). The base is also the departure point for surveillance flights over Gaza conducted by the UK, which started in December 2023, though the UK Government has denied that this information is being used for the purposes of targeting (Kennard, 2024).

Whether Israel has access to the data of the major commercial internet companies is less clear. A “target machine” proposed in a 2021 book by Brigadier General Yossi Sariel, the head of the IDF Cyberwarfare Unit 8200, analyses an individual's risk based on factors including their social connections and their membership of chat groups with individuals who are already deemed suspect. On the basis of this, it has been speculated that the IDF may have (or at least want) access to WhatsApp metadata (Biggar, 2024).

There are a number of points to make about Lavender and its related systems in the context of IHL and debates on regulation of AWS. Firstly, it is important to understand that this kind of automated targeting system does *not* constitute an AWS according to the ICRC definition; while an automated system suggests targets, the strike itself is still ordered by an individual.

This is not to say that Israel is not a player in AWS development. In 2019 Stuart Russell saw Israel's Harop loitering munition as “probably the clearest existing example of LAWs... It searches for up to six hours in a given geographical region for any target that meets a given criterion and then destroys it” (Russell, 2021). The IDF has also claimed to have employed drone swarms in Operation Guardian of the Walls (Gross, 2021).

Internationally, Israel has been opposed to all attempts at international governance over the development and use of LAWs. It was one of very few countries not to support the near-universal UN General Assembly vote of November 2023, and has also not signed up to the US Political Declaration (Automated Decision Research, 2024).

Israel is increasingly important as an exporter of military equipment, particularly in the realm of surveillance technology where it has been estimated to be the leading exporter worldwide in terms of income per capita. Much of the technology forming

Israel's export offer has been "battle-tested" in Gaza and the Occupied Palestinian territories (Loewenstein, 2023; Shehabi and Lubin, 2024).

Israel's AI-enabled target development, and the decisions of military policy that surround it, pose clear problems for the foundational IHL principles of distinguishing between combatants and civilians, military necessity, proportionality and the duty to prevent errors – not to mention accountability for abiding by all of these.

To take the duty to distinguish between civilians and combatants first, the logic of targeting described by +972 and *Local Call* is one where any male of adult age - and reportedly also those under the age of 17 - is vulnerable to being named as a target. A probabilistic assessment of an individual's proximity to the governing party in Gaza could easily take in a range of civilian workers. Indeed at least one of +972's sources claimed that this conflation was present in the training data.

In previous conflicts conducted by the IDF, where the number of targets was significantly lower, the identities of human targets were verified by human intelligence analysis. Now, the absence of this kind of threshold meant that there was no logical end to the list produced by Lavender and related tools. "Because of the system, the targets never end," stated a source for +972.

It is also clear that the targeting policy in Gaza is characterised by a very high tolerance of civilian casualties and therefore a lack of proportionality. Sources for +972 and *Local Call* stated that the number of civilians likely to be in a given location was known prior to any strike. Where's Daddy was developed in order to facilitate attacks on family residences, which happened largely during nighttime hours.

The decision about the number of acceptable civilian deaths per targeted person varied over the course of the conflict, but could be between 15 and 20. For "high priority" targets, the number of acceptable civilian deaths has been much higher, sometimes 100 or more.

According to one source quoted:

"Nothing happens by accident... When a 3-year-old girl is killed in a home in Gaza, it's because someone in the army decided it wasn't a big deal for her to be killed — that it was a price worth paying in order to hit [another] target."

Indeed, the use of 'power targets', including entire residential blocks, was part of a concerted strategy whereby the sheer scale of damage would potentially generate pressure against Hamas from the Palestinian population - all of which is clearly problematic from the perspective of military necessity.

Even within the logic of the system, the error rate associated with identified targets was in the region of 10% - meaning that one in ten individuals appearing on the IDF's target list were not members of either of the two designated groups. Latency in the intelligence used in Where's Daddy meant that targets could also be elsewhere at the time a strike was initiated.

All of which, of course, makes the role of a human in the loop checking system outputs all the more important, both for limiting error and ensuring some degree of accountability. However, the role of military personnel in overseeing the target lists produced by Lavender is, according to +972 and *Local Call*, extremely cursory. In the initial stages of the war, soldiers were instructed to trust the outputs of the system to a large extent, in the interests of processing as many targets as possible.

One source told +972 and *Local Call* that "I had zero added value as a human, apart from being a stamp of approval." Another described spending a maximum of "twenty seconds" reviewing target lists, usually solely to check whether individuals on the list were male, which was done by listening to an audio recording of their voice. There was no real mechanism for detecting or correcting where male civilians had been added to the target list, for instance when phones were shared between friends or family members.

Operators were apparently aware of the imprecision of the strikes being conducted and the high number of civilian casualties. Sources for reporting stated that the presence of a 'neutral' technological system made it easier psychologically to accept their role in such strikes. "The machine did it coldly. And that made it easier," one source told +972. This bears some resemblance to the "technification" dynamic described by Schwartz and Remic, reducing the moral agency of both the targeted and the human on the operating end (Renic and Schwartz, 2023).

Many of those speaking to +972 about these systems were aware of the shortcuts being taken. Given the current state of public opinion in Israel, speaking to the media can itself be seen as a form of resistance. It may be significant that some of the most visible collective action on military policy, itself modest, has come from reservists who have been called up to serve in the present conflict and are very aware of the changes that have taken place since their earlier compulsory military service (Goodfriend, 2024).

All this recalls Paul Lushenko's conclusions about military personnels' support for AI-enabled systems; the overall context is important and the impression of being involved in a 'justified' conflict could well allow misgivings to be overcome. "US officers may feel obliged to embrace projected forms of warfare that go against their own preferences and attitudes". In a context where domestic opposition to an ongoing conflict has been marginalised, it is easy to see how human in the loop mechanisms can become ever more ineffectual (Lushenko, 2023; Rapoport, 2024).

Implications for limiting AWS

Among commentators, there is a virtual consensus that this AI-enabled targeting system does not conform with IHL precepts. This opinion is shared by observers who are generally supportive of US military policy in the digital realm as well as international agencies (Feldstein, 2024). The UN AI Advisory Body's final report *Governing for Humanity*, published in September 2024, is a case in point:

"The Advisory Body considers it essential to identify clear red lines delineating unlawful use cases, including relying on AI to select and engage targets autonomously" (UN AI Advisory Body, 2024).

Lavender poses profound problems for the formulation of minimal standards in military domain AI that promise compatibility with IHL, particularly in terms of testing before deployment and the post-facto assessment of what happens when these systems are used on the battlefield. Much of this was, of course, foreseen by the military theorists who have considered this issue.

The recognition that the battlefield has its own logic, tending towards swift deployment of new technology for competitive advantage, essentially "moving fast and breaking things," is inimical to calls for testing and audits, as well as deep human involvement in the control of systems at the design stage. We have seen similar dynamics in the rapid prototyping and deployment of low-cost drones in Ukraine (Hudson and Khudov, 2023).

As *+972* and *Local Call* reported, post strike assessments, checking whether targets had actually been present, have been largely abandoned after October 7.

While these critiques should of course be seen in the immediate humanitarian context of a military operation that has, at the time of writing, ended tens of thousands of lives and destroyed 90% of Gaza's housing stock and the majority of its civic infrastructure, not to mention one which the ICJ has found could plausibly amount to genocide, there are additional reasons for commentators' concerns.

The revelations about Lavender have been seen as disastrous from the perspective of limiting the development and proliferation of lethal AI-enabled technology. As noted above, there is a broad awareness that proposals on limiting AWS, which have been the focus of arms control community in this area for many years, will not be sufficient to meet this challenge.

There is also a widely expressed concern that the US – potentially the only geopolitical force in a position to establish limits on Israel's operations – has failed to do so. Lavender and its accompanying systems breach the limited standards proposed in the US Political Declaration, which aim at bare compliance with IHL. Indeed, as we have seen, the US is not only refraining from criticising Israeli targeting policy, it is likely also directly assisting it by sharing satellite footage and other intelligence gathered using surveillance drones.

Reliance on civilian technology

Notwithstanding speculation about the IDF's ability to access metadata produced by the major internet companies, there is no question that Lavender and other AI-enabled applications used in Gaza rely on the capabilities of Big Tech. This is true in several dimensions.

Firstly, providing services to militaries is likely to become increasingly important to the major AI labs. Despite the growing ubiquity of ChatGPT, OpenAI is at present not a profitable company. In early 2024, OpenAI quietly abandoned its commitment not

to produce military applications and has since concluded several agreements with the US Department of Defense (Field, 2024; Biddle, 2024).

More than that, if the factors critical to the development of AI systems are considered to be the triad of algorithms, data and computing power ('compute'), while it is conceivable that nation states may be able to draw on data and algorithms independently, very few are able to draw on sufficient native computing resources ('sovereign compute') in the absence of services provided by commercial providers.¹² Observers have described "the emergence of a new military industrial complex that is dependent on commercial service providers" (Lushenko and Carter, 2024).

This is certainly true of IDF operations in Gaza, where Google, Microsoft and Amazon Web Services (AWS) had contracts with Israel prior to October 2023. Project Nimbus, a joint contract held by Google and Amazon since 2021 to provide the Israeli government with cloud computing, AI and other services, was "deepened" as Israel's computing requirements expanded. According to a draft contract seen by journalists, services are provided to Israel's Ministry of Defence directly. IDF personnel interviewed by +972 and *Local Call* refer to using infrastructure from all three companies (Osseiran and Asher-Schapiro, 2024; Perrigo, 2024). Reports suggest that AWS servers are used to store mass surveillance data, particularly audio recordings, with Google and Microsoft AI tools being used to process it (Abraham, 2024a).

While the major tech companies have not been responsive to calls from civil society to observe their human rights obligations, these revelations have led to a number of employee protests at the major tech companies (Fatafa, 2024). Project Nimbus had, in fact, been controversial among Google and Amazon employees since its inception, with an open letter signed by hundreds of employees from both companies in 2021 (Anonymous Google and Amazon workers, 2021).

In April 2024, shortly after the publication of details from the new draft Project Nimbus contract, Google fired at least 50 employees who had staged sit-ins at company offices. Police were even brought in to arrest nine Google employees. Other tech workers at Google, some in prominent positions, resigned (Del Valle, 2024; Heath, 2024). At least two workers were also fired at Microsoft in the autumn of 2024 (O'Brien, 2024).

Conclusion

Discussions about limiting the development and proliferation of AWS through new, binding international rules started over a decade ago and, at the time of writing, are due to be progressed in the current session of the UN General Assembly. The conclusion of a new treaty is supported by the majority of UN member states.

¹² Both the US and EU, and to some extent the UK, are investing heavily to develop these capabilities at present.

However, given developments on the ground in the world's war zones in 2024, the limitations imposed by the slowness of international negotiations are readily apparent: we may already have missed the chance to prevent AI-enabled warfare causing new levels of automated human suffering and destabilising the international order generally.

Sceptics of the value of AI for military purposes have counselled caution about new legal rules on the grounds of the technology's reliance on comprehensive data, its probable primary utility for intelligence and logistics functions, its unpredictability and the likely incompatibility of "meaningful human control" with the speed that is the technology's foremost advantage for militaries. These are, of course, exactly the conditions we have seen in the operation of automated targeting in Gaza.

Given the failure to enforce even weaker, non-binding controls, such as those in the US-led Political Declaration on Responsible Military use of Artificial Intelligence and Autonomy, there is an urgent need to think about alternative approaches. One recent proposal notes the dominance of the civilian sector in the development of AI and the likely future reliance of militaries, and the public sector in general, on the frontier models developed by the major AI labs. As such, the authors suggest that a focus on excluding personal identifying information from training data in general could mitigate the dangers of militaries using AI to target individuals (Khlaaf, West and Whitaker, 2024).

The increasing dependency of military on civilian technology may yet present a way forward. Perhaps surprisingly, and unlike their counterparts (and major investors) in Big Tech, the major labs have proven themselves receptive to public sector intervention, at least to some extent. The advance of AI Safety has seen the major labs welcoming some degree of external scrutiny, to the extent of making formal agreements for access to commercially sensitive internal data. There are, of course, major advantages for them in doing so, not least in terms of sharing risk and offloading a certain degree of financial investment in safety research and auditing.

The UK has not been a progressive force in regulating AWS. Not only did the last government definitively exempt itself from new rules-making in this area in response to the House of Lords Special Inquiry, the UK has actively tried to obstruct any agreement by blocking agreement in the CCW. In light of the UK's focus on AI safety in the civilian domain, which has seen it make a significant financial and diplomatic investment in developing methodology and criteria for audit internationally, its position on safety and accountability on military AI, which relies on the difficult-to-define concept of "meaningful human control," now looks highly anomalous. At present, as the 2023 House of Lords Special Inquiry found, the two domains have been kept strictly separate (AI Committee, 2023, para. 88-89). This division is not only undesirable; in the medium-term, it is unlikely to be sustainable.

References

- Abraham, Y. (2023) "A mass assassination factory": Inside Israel's calculated bombing of Gaza', +972 Magazine, 30 November. Available at: <https://www.972mag.com/mass-assassination-factory-israel-calculated-bombing-gaza/> (Accessed: 18 October 2024)
- Abraham, Y. (2024) "'Lavender': The AI machine directing Israel's bombing spree in Gaza", +972 Magazine, 3 April. Available at: <https://www.972mag.com/lavender-ai-israeli-army-gaza/> (Accessed: 18 October 2024)
- Abraham, Y. (2024a) "Order from Amazon": How tech giants are storing mass data for Israel's war', +972 Magazine, 4 August. Available at: <https://www.972mag.com/cloud-israeli-army-gaza-amazon-google-microsoft/> (Accessed: 18 October 2024)
- Ada Lovelace Institute and the Alan Turing Institute. (2023) How do people feel about AI? A nationally representative survey of public attitudes to artificial intelligence in Britain. Available at: <https://adalovelaceinstitute.org/report/public-attitudes-ai> (Accessed: 18 October 2024)
- Ahronheim, A. (2021) 'Israel's operation against Hamas was the world's first AI war', The Jerusalem Post, 27 May. Available at: <https://www.jpost.com/arab-israeli-conflict/gaza-news/guardian-of-the-walls-the-first-ai-war-669371> (Accessed: 18 October 2024)
- AI Safety Institute. (2024) 'About.' Available at: <https://www.aisi.gov.uk/about> (Accessed: 18 October 2024)
- Al Shabaka. (2024) 'Gaza: Testing Ground for Israel's Global Weapons Industry', 29 August. (online video). Available at: <https://www.youtube.com/watch?v=lgp87-jel2U> (Accessed 18 October 2024)
- Allen, J.R. and Husain, A. (2017) 'On Hyperwar', Proceedings Vol. 143/7/1,373, US Naval Institute, June. Available at: <https://www.usni.org/magazines/proceedings/2017/july/hyperwar> (Accessed: 18 October 2024)
- Amnesty International. (2023) 'Israel using previously-unreported facial recognition system to 'automate apartheid' against Palestinians – new report', 2 May. Available at: <https://www.amnesty.org.uk/press-releases/israel-using-previously-unreported-facial-recognition-system-automate-apartheid> (Accessed: 18 October 2024)
- Anderljung, M., Smith, E.T., O'Brien, J., Soder, L., Bucknall, B., Bluemke, E., Schuett, J., Trager, R., Strahm, L. and Chowdhury, R. (2023) 'Towards publicly accountable frontier LLMs: Building an external scrutiny ecosystem under the ASPIRE framework', 15 November. Available at: <https://arxiv.org/abs/2311.14711> (Accessed: 18 October 2024)
- Anonymous Google and Amazon workers. (2021) 'We are Google and Amazon workers. We condemn Project Nimbus', The Guardian, 12 October. Available at: <https://www.theguardian.com/commentisfree/2021/oct/12/google-amazon-workers-condemn-project-nimbus-israeli-military-contract> (Accessed: 18 October 2024)
- Automated Decision Research. (2024) 'States Positions.' Available at: <https://automatedresearch.org/state-positions/> (Accessed: 18 October 2024)
- Automated Decision Research. (2024a) 'Weapons Systems.' Available at: <https://automatedresearch.org/weapons-systems/> (Accessed: 18 October 2024)
- Biddle, S. (2024) 'US Military Makes First Confirmed OpenAI Purchase for War Fighting Forces', The Intercept, 25 October. Available at: <https://theintercept.com/2024/10/25/africom-microsoft-openai-military/> (Accessed: 28 October 2024)
- Biggar, P. (2024) 'Meta and Lavender', 16 April. Available at: <https://blog.paulbiggar.com/meta-and-lavender/#:~:text=Notes,-%5B1%5D,-From%20the%20%2B972> (Accessed: 18 October 2024)
- BlueDot Impact. (2024) 'AI Governance Course', AI Safety Fundamentals. Available at: <https://course.aisafetyfundamentals.com/governance> (Accessed: 18 October 2024)
- Bostrom, N. (2014) Superintelligence: Paths, Dangers, Strategies. Oxford: Oxford University Press.

- Buchholz, K. (2024) 'The world's biggest arms exporters',
- Chakravorti, B., Bhalla, A. and Chaturvedi, R. S. (2023) 'Charting the Emerging Geography of AI', *Harvard Business Review*, 12 December. Available at: <https://hbr.org/2023/12/charting-the-emerging-geography-of-ai> (Accessed: 18 October 2024)
- Chopra, C., Kasare, A. and Gupta, P. (2024) 'How venture capital is investing in AI in the top five global economies – and shaping the ecosystem', *World Economic Forum*, 24 May. Available at: <https://www.weforum.org/agenda/2024/05/these-5-countries-are-leading-the-global-ai-race-heres-how-theyre-doing-it/> (Accessed: 18 October 2024)
- Clarke, L. (2023) 'How Silicon Valley doomers are shaping Rishi Sunak's AI plans', *Politico*, 14 September. Available at: <https://www.politico.eu/article/rishi-sunak-artificial-intelligence-pivot-safety-summit-united-kingdom-silicon-valley-effective-altruism/> (Accessed: 18 October 2024)
- Crary, A. (2023) 'The toxic ideology of longtermism', *Radical Philosophy* 214, April 2023, pp. 49–57. Available at: https://www.radicalphilosophy.com/wp-content/uploads/2023/04/rp214_crary.pdf (Accessed: 18 October 2024)
- Danzig, R. (2018) 'Technology Roulette: Managing Loss of control as Many Militaries Pursue Technological Superiority', *Center for a New American Century*, June. Available at: <https://s3.amazonaws.com/files.cnas.org/documents/CNASReport-Technology-Roulette-DoSproof2v2.pdf?mtime=20180628072101> (Accessed: 18 October 2024)
- deCoriolis, A., Gross, A.S., Tuminello, J. Gross, S.J. and Channin, J. (2023) 'Animal Advocacy's Stockholm Syndrome', in *The Good It Promises, the Harm It Does*, edited by Adams, C.J., Crary, A., and Gruen, L. Oxford: Oxford University Press.
- Defence, Science and Technology Laboratory (DSTL). (2024) AUKUS takes another step forward with real-time AI trials, 9 August. Available at: <https://www.gov.uk/government/news/aukus-takes-another-step-forward-with-real-time-ai-trials> (Accessed: 18 October 2024)
- Del Valle, G. (2024) 'Nine Google employees arrested after eight-hour sit-in protest', *The Verge*, 17 April. Available at: <https://www.theverge.com/2024/4/17/24133056/google-protests-project-nimbus-no-tech-for-apartheid> (Accessed: 18 October 2024)
- Department for Science, Innovation and Technology (DSIT) and Office for Artificial Intelligence. (2023) 'AI regulation: a pro-innovation approach', 29 March. Available at: <https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach> (Accessed: 18 October 2024)
- Department for Science, Innovation and Technology (DSIT). (2023) 'Tech entrepreneur Ian Hogarth to lead UK's AI Foundation Model Taskforce', 18 June. Available at: <https://www.gov.uk/government/news/tech-entrepreneur-ian-hogarth-to-lead-uks-ai-foundation-model-taskforce> (Accessed: 18 October 2024)
- Department for Science, Innovation and Technology (DSIT). (2023a) 'Industry and national security heavyweights to power UK's Frontier AI Taskforce', 7 September. Available at: <https://www.gov.uk/government/news/industry-and-national-security-heavyweights-to-power-uks-frontier-ai-taskforce> (Accessed: 18 October 2024)
- Department for Science, Innovation and Technology (DSIT). (2024) 'Government's trailblazing Institute for AI Safety to open doors in San Francisco', 20 May. Available at: <https://www.gov.uk/government/news/governments-trailblazing-institute-for-ai-safety-to-open-doors-in-san-francisco> (Accessed: 18 October 2024)
- Ding, J. and Dafoe, A. (2023) 'Engines of power: Electricity, AI, and general-purpose, military transformations', *European Journal of International Security*, 8(3), pp. 377–394. doi:10.1017/eis.2023.1.
- Dumbi, A. R. (2023) 'Commander of the Artificial Intelligence Center 8200: Artificial Intelligence enables faster classification and detection of terrorist targets', *Israel Defense*, 14 February. Available at: <https://www.israeldefense.co.il/node/57256> (Accessed: 18 October 2024)

- Dwoskin, E. (2021) 'Israel escalates surveillance of Palestinians with facial recognition program in West Bank', Washington Post, 8 November. Available at: https://www.washingtonpost.com/world/middle_east/israel-palestinians-surveillance-facial-recognition/2021/11/05/3787bf42-26b2-11ec-8739-5cb6aba30a30_story.html (Accessed: 18 October 2024)
- Edwards, L. and Veale, M. (2017) 'Slave to the Algorithm? Why a 'Right to an Explanation' Is Probably Not the Remedy You Are Looking For,' 16 *Duke Law & Technology Review* 18-84 (2017)
- Emerson, S. and Nieva, R. (2024) 'Eric Schmidt is secretly testing AI military drones in a wealthy Silicon Valley suburb', Forbes, 6 June. Available at: <https://www.forbes.com/sites/sarahemerson/2024/06/06/eric-schmidt-is-secretly-testing-ai-military-drones-in-a-wealthy-silicon-valley-suburb/> (Accessed: 18 October 2024)
- Farah, H. (2023) 'Israeli restrictions leave Palestinians facing digital divide', The Guardian, 30 October. Available at: <https://www.theguardian.com/world/2023/oct/30/israeli-restrictions-leave-palestinians-facing-digital-divide> (Accessed: 18 October 2024)
- Fatafa, M. (2024) 'Big Tech and the risk of genocide in Gaza: what are companies doing?', Access Now, 10 October. Available at: <https://www.accessnow.org/gaza-genocide-big-tech/> (Accessed: 18 October 2024)
- Feldstein, S. (2024) 'AI in war: Can advanced military technologies be tamed before it's too late?', Bulletin of the Atomic Scientists, 11 January. Available at: <https://thebulletin.org/2024/01/ai-in-war-can-advanced-military-technologies-be-tamed-before-its-too-late/> (Accessed: 18 October 2024)
- Field, H. (2022) 'Google, Amazon face new protests over contract with Israeli government and military', Emerging Tech Brew, 19 September. Available at: <https://www.emergingtechbrew.com/stories/2022/09/19/google-amazon-face-new-protests-over-contract-with-israeli-government-and-military> (Accessed: 18 October 2024)
- Field, H. (2024) 'OpenAI quietly removes ban on military use of its AI tools', CNBC, 16 January. Available at: <https://www.cnbc.com/2024/01/16/openai-quietly-removes-ban-on-military-use-of-its-ai-tools.html> (Accessed: 18 October 2024)
- Future of Life Institute. (2016) 'Autonomous Weapons Open Letter: AI & Robotics Researchers', 9 February. Available at: <https://futureoflife.org/open-letter/open-letter-autonomous-weapons-ai-robotics/?cn-reloaded=1> (Accessed: 18 October 2024)
- Gebru, T., & Torres, Émile P. (2024). The TESCREAL bundle: Eugenics and the promise of utopia through artificial general intelligence. First Monday, 29(4). <https://doi.org/10.5210/fm.v29i4.13636>
- Goldfarb, A. and Lindsay, J.R. (2022) 'Prediction and Judgment: Why Artificial Intelligence Increases the Importance of Humans in War', International Security, 46 (3): 7-50. Available at: https://doi.org/10.1162/isec_a_00425 (Accessed: 18 October 2024)
- Goodfriend, S. (2024) 'Kill Lists', London Review of Books, 10 October. Available at: <https://www.lrb.co.uk/the-paper/v46/n19/sophia-goodfriend/kill-lists> (Accessed: 18 October 2024)
- Gross, A. and Thomas, D. (2024) 'UK to consult on 'opt-out' model for content scraping in blow to publishers', Financial Times, 16 October. Available at: <https://www.ft.com/content/26bc3de1-af90-4c69-9f53-61814514aeaa> (Accessed: 18 October 2024)
- Gross, J.A. (2021) 'In apparent world first, IDF deployed drone swarms in Gaza fighting', Times of Israel, 10 July. Available at: <https://www.timesofisrael.com/in-apparent-world-first-idf-deployed-drone-swarms-in-gaza-fighting/> (Accessed: 18 October 2024)
- Halteman, M.C. (2023) 'Diversifying Effective Altruism's Long Shots in Animal Advocacy', in The Good It Promises, the Harm It Does, edited by Adams, C.J., Crary, A., and Gruen, L. Oxford: Oxford University Press.
- Heath, A. (2024) 'Google fires 28 employees after sit-in protest over Israel cloud contract', The Verge, 18 April. Available at: <https://www.theverge.com/2024/4/17/24133700/google-fires-28-employees-protest-israel-cloud-contract> (Accessed: 18 October 2024)

- HL Deb (19 April 2024). vol. 837, col. 1230. Available at: <https://hansard.parliament.uk/Lords/2024-04-19/debates/278564E2-0E85-4CBE-BACA-953EC93B93E4/ArtificialIntelligenceInWeaponSystemsCommitteeReport#contribution-9DB3CE28-B2C2-4A03-8CF7-E16FC469862A> (Accessed: 18 October 2024)
- HL Deb (19 April 2024a). vol. 837, col. 1248. Available at: <https://hansard.parliament.uk/Lords/2024-04-19/debates/278564E2-0E85-4CBE-BACA-953EC93B93E4/ArtificialIntelligenceInWeaponSystemsCommitteeReport#contribution-9E79499C-2D3E-471B-80B7-2C9DDA3B6301> (Accessed: 18 October 2024)
- Hogarth, I. (2023) 'We must slow down the race to god-like AI', Financial Times, 13 April. Available at: <https://www.ft.com/content/03895dc4-a3b7-481e-95cc-336a524f2ac2> (Accessed 18 October 2024)
- House of Lords AI in Weapons Systems Committee (AI Committee). (2023a) 'Proceed with Caution: Artificial Intelligence in Weapons Systems', 1 December. Available at: <https://committees.parliament.uk/publications/42387/documents/210740/default/> (Accessed 18 October 2024)
- House of Lords Communications and Digital Committee (Communications and Digital Committee). (2023) Corrected oral evidence: large language models, 12 September. Available at: <https://committees.parliament.uk/oralevidence/13655/html/> (Accessed 18 October 2024)
- House of Lords Communications and Digital Committee (Communications and Digital Committee). (2023a) Corrected oral evidence: large language models, 17 October. Available at: <https://committees.parliament.uk/oralevidence/13692/html/> (Accessed 18 October 2024)
- Hudson, J. and Khudov, K. (2023) 'The war in Ukraine is spurring a revolution in drone warfare using AI', Washington Post, 26 July. Available at: <https://www.washingtonpost.com/world/2023/07/26/drones-ai-ukraine-war-innovation/> (Accessed 18 October 2024)
- Human Rights Watch. (2024) 'Questions and Answers: Israeli Military's Use of Digital Tools in Gaza', 10 September. Available at: <https://www.hrw.org/news/2024/09/10/questions-and-answers-israeli-militarys-use-digital-tools-gaza> (Accessed 18 October 2024)
- International Committee of the Red Cross (ICRC). (2021) 'ICRC position on autonomous weapons systems', 12 May. Available at: <https://www.icrc.org/en/document/icrc-position-autonomous-weapon-systems> (Accessed 18 October 2024)
- International Committee of the Red Cross (ICRC). (2023) 'Joint call by the United Nations Secretary-General and the President of the International Committee of the Red Cross for States to establish new prohibitions and restrictions on Autonomous Weapon Systems', 5 October. Available at: <https://www.icrc.org/en/document/joint-call-un-and-icrc-establish-prohibitions-and-restrictions-autonomous-weapons-systems> (Accessed 18 October 2024)
- International Committee of the Red Cross (ICRC). (2024) 'Targeting under International Humanitarian Law', How does law protect in war? Available at: <https://casebook.icrc.org/highlight/targeting-under-international-humanitarian-law> (Accessed 18 October 2024)
- Kennard, M. (2024) 'Kier Starmer's 100 spy flights over Gaza in support of Israel', Declassified UK, 3 October. Available at: <https://www.declassifieduk.org/keir-starmer-100-spy-flights-over-gaza-in-support-of-israel/> (Accessed: 18 October 2024)
- Kennard, M. (2024a) 'US Special Ops flights to Israel from UK's Cyprus base surge under Starmer', Declassified UK, 10 October. Available at: <https://www.declassifieduk.org/u-s-special-ops-flights-to-israel-from-uks-cyprus-base-surge-under-starmer/> (Accessed: 18 October 2024)
- Khlaaf, H., West, S.M. and Whittaker, M. (2024) 'Mind the Gap: Foundation Models and the Covert Proliferation of Military Intelligence, Surveillance, and Targeting'. *arXiv preprint arXiv:2410.14831*.
- Klippenstein, K. and Petti, M. (2024) 'Biden Admin deployed Air Force team to Israel to assist with targets, document suggests', The Intercept, 11 January. Available at: <https://theintercept.com/2024/01/11/israel-air-force-targeting-intelligence/> (Accessed: 18 October 2024)

- Kototajlo, D. and Ball, D.W. (2024) '4 ways to advance transparency in frontier AI development', Time, 15 October. Available at: <https://time.com/collection/time100-voices/7086285/ai-transparency-measures/> (Accessed: 18 October 2024)
- Lamparth, M., Corso, A., Ganz, J., Mastro, O.S., Schneider, J. and Trinkunas, H., (2024). Human vs. machine: Language models and wargames. *arXiv preprint arXiv:2403.03407*.
- Loewenstein, A. (2023) The Palestine Laboratory: How Israel Exports the Technology of Occupation Around the World. Verso, London.
- Lushenko, P. (2023) 'AI and the future of warfare', Bulletin of the Atomic Scientists, 29 November. Available at: <https://thebulletin.org/2023/11/ai-and-the-future-of-warfare-the-troubling-evidence-from-the-us-military/> (Accessed: 18 October 2024)
- Luskenko, P. and Carter, K. (2024) 'A new military-industrial complex', Bulletin of the Atomic Scientists, 7 October. Available at: <https://thebulletin.org/2024/10/a-new-military-industrial-complex-how-tech-bros-are-hyping-ais-role-in-war/> (Accessed: 18 October 2024)
- McBride, K. (2023) 'Why the UK AI Safety Summit will fail to be meaningful', Oxford Internet Institute, 29 September. Available at: <https://www.oii.ox.ac.uk/news-events/dr-keegan-mcbride-why-the-uk-ai-safety-summit-will-fail-to-be-meaningful/> (Accessed: 18 October 2024)
- McQuillan, D. (2022) Resisting AI: An Anti-Fascist Guide to Artificial Intelligence. Bristol: Bristol University Press
- Ministry of Defence (MoD). (2022) June. Defence Artificial Intelligence Strategy. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1082416/Defence_Artificial_Intelligence_Strategy.pdf (Accessed: 18 October 2024)
- Ministry of Defence (MoD). (2022a). Ambitious, safe, responsible: our approach to the delivery of AI-enabled capability in Defence, 15 June. Available at: <https://www.gov.uk/government/publications/ambitious-safe-responsible-our-approach-to-the-delivery-of-ai-enabled-capability-in-defence/ambitious-safe-responsible-our-approach-to-the-delivery-of-ai-enabled-capability-in-defence> (Accessed: 18 October 2024)
- Ministry of Defence (MoD). (2023) British Army Approach to Artificial Intelligence, 1 October. Available at: https://www.army.mod.uk/media/24745/20231001-british_army_approach_to_artificial_intelligence.pdf (Accessed: 18 October 2024)
- Ministry of Defence (MoD) AI Ethics Advisory Panel. (2024) Minute, 25 January. Available at: https://assets.publishing.service.gov.uk/media/662243cdf49bd20327a0d788/2024-01-25_MOD_AI_Ethics_Advisory_Panel_Minutes.pdf (Accessed: 18 October 2024)
- National Institute of Standards and Technology (NIST). (2024) 'U.S. AI Safety Institute signs agreements regarding AI safety research, testing and evaluation with Anthropic and OpenAI', 29 August. Available at: <https://www.nist.gov/news-events/news/2024/08/us-ai-safety-institute-signs-agreements-regarding-ai-safety-research> (Accessed: 18 October 2024)
- O'Brien, M. (2024) Microsoft fires employees who organized vigil for Palestinians killed in Gaza 26 October. Available at: <https://apnews.com/article/microsoft-fired-workers-israel-palestinians-gaza-72de6fe1f35db9398e3b6785203c6bbf> (Accessed: 18 October 2024)
- Osseiran, N. and Asher-Schapiro, A. (2024) 'One Year of War in Gaza: Decoding the role of Big Tech', Context, 7 October. Available at: <https://www.context.news/big-tech/one-year-of-war-in-gaza-decoding-the-role-of-big-tech> (Accessed: 18 October 2024)
- Perrigo, B. (2024) 'Exclusive: Google Contract Shows Deal With Israel Defense Ministry', Time, 12 April. Available at: <https://time.com/6966102/google-contract-israel-defense-ministry-gaza-war/> (Accessed: 18 October 2024)
- Petroupoulos, A. (2024) 'The AI Safety Institute Network: who, what and how?', ICFG, September. Available at: <https://icfg.eu/the-ai-safety-institute-network-who-what-and-how/> (Accessed: 18 October 2024)

- Rapoport, M. (2024) 'Israeli society's dehumanization of Palestinians is now absolute', +972 Magazine, 23 August. Available at: <https://www.972mag.com/dehumanization-moral-abyss-israelis/> (Accessed: 18 October 2024)
- Renic, N. and Schwarz, E. (2023) 'Crimes of Dispassion: Autonomous Weapons and the Moral Challenge of Systematic Killing', *Ethics & International Affairs*, 37(3), pp. 321–343. doi:10.1017/S0892679423000291.
- Right to Warn. (2024) A right to warn about advanced artificial intelligence, 4 June. Available at: <https://righttowarn.ai> (Accessed 18 October 2024)
- Rowe, N. (2023) "It's destroyed me completely": Kenyan moderators decry toll of training of AI models', The Guardian, 2 August. Available at: <https://www.theguardian.com/technology/2023/aug/02/ai-chatbot-training-human-toll-content-moderator-meta-openai> (Accessed 18 October 2024)
- Russell, S. (2019) *Human Compatible: AI and the problem of control*. London: Penguin.
- Saltini, A. (2023) 'To avoid nuclear instability, a moratorium on integrating AI into nuclear decision-making is urgently needed: The NPT PrepCom can serve as a springboard', European Leadership Network. Available at: <https://www.europeanleadershipnetwork.org/commentary/to-avoid-nuclear-instability-a-moratorium-on-integrating-ai-into-nuclear-decision-making-is-urgently-needed-the-npt-prepcom-can-serve-as-a-springboard/> (Accessed 18 October 2024)
- Scharre, P. (2019) *Army of None: Autonomous Weapons and the Future of War*. New York: Norton.
- Scharre, P. (2023) *Four Battlegrounds: Power in the age of Artificial Intelligence*. New York: Norton.
- Sellman, M. (2023) 'Tony Blair and William Hague: AI revolution could leave Britain behind', The Times, 13 June. Available at: <https://www.thetimes.co.uk/article/tony-blair-and-william-hague-ai-revolution-could-leave-britain-behind-278r0fz2x> (Accessed: 18 October 2024)
- Shehabi, O.M. and Lubin, A. (2024) 'Israel-Hamas 2024 Symposium – Algorithms of war: Military AI and the war in Gaza', Lieber Institute West Point, 24 January. Available at: <https://lieber.westpoint.edu/algorithms-war-military-ai-war-gaza/> (Accessed: 18 October 2024)
- Skelton, S. (2023) 'UK, US and Australia jointly trial AI-enabled drone swarm', Computer Weekly, 30 May. Available at: <https://www.computerweekly.com/news/366538738/UK-US-and-Australia-jointly-trial-AI-enabled-drone-swarm> (Accessed: 18 October 2024)
- Sunak, R. (2023) 'Rishi Sunak & Elon Musk: Talk AI, Tech & the Future' (online video). Available at: <https://www.youtube.com/watch?v=R2meHtrO1n8> (Accessed 18 October 2024)
- Timberg, C. and Soltani, A. (2013) 'By cracking cellphone code, NSA has ability to decode private conversations', Washington Post, 13 December. Available at: https://www.washingtonpost.com/business/technology/by-cracking-cellphone-code-nsa-has-capacity-for-decoding-private-conversations/2013/12/13/e119b598-612f-11e3-bf45-61f69f54fc5f_story.html (Accessed: 18 October 2024)
- Tony Blair Institute for Global Change. (2023) A new national purpose: AI promises a world-leading future for Britain, 13 June. Available at: <https://institute.global/insights/politics-and-governance/new-national-purpose-ai-promises-world-leading-future-of-britain> (Accessed: 18 October 2024)
- UK Campaign to Stop Killer Robots. (2022) Stop Killer Robots in UK Universities, September. Available at: https://una.org.uk/sites/default/files/2022-09/stop_killer_robots_in_uk_universities_-_report_and_annex_-_20_sept_2022.pdf (Accessed: 18 October 2024)
- UK Parliament. (2024) Written question on Israel: Military Aid, 5 December, UIN 4501. Available at: <https://questions-statements.parliament.uk/written-questions/detail/2024-09-05/4501/> (Accessed: 18 October 2024)
- United Nations AI Advisory Body. (2024) Governing AI for Humanity. September. Available at: https://www.un.org/sites/un2.un.org/files/governing_ai_for_humanity_final_report_en.pdf (Accessed: 18 October 2024)

United Nations Institute for Disarmament Research (UNIDIR) (2023). Artificial Intelligence beyond weapons, October. Available at: https://unidir.org/wp-content/uploads/2023/10/UNIDIR_AI_Beyond_Weapons_Application_Impact_AI_in_the_Military_Domain.pdf (Accessed: 18 October 2024)

US Department of State. (2023) 'Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy', Bureau of Arms Control, Deterrence and Stability, 9 November. Available at: <https://www.state.gov/political-declaration-on-responsible-military-use-of-artificial-intelligence-and-autonomy-2/> (Accessed: 18 October 2024)