12 AlphaGo's Move 37 and Its Implications for AI-Supported Military Decision-Making

Thomas W. Simpson

INTRODUCTION

The most dramatic use-case for Artificial Intelligence (AI) in military contexts is weapons systems that have the capacity to identify targets, including human combatants, and engage them autonomously. Imagine a drone with image recognition technology, able to identify enemy combatants by their uniform and authorized to target them with a rifle without any human involvement in either target identification or engagement. Such a system meets the US Department of Defense's Directive 3000.09 definition of an 'autonomous weapon system' in that, 'once activated, [it] can select and engage targets without further intervention by an operator' (2023, p. 21), and colloquially may be described as having humans 'off the loop'. Given the severe consequences of such a system malfunctioning and killing people whom the laws of war protect from lethal force—including civilians, enemy combatants who are surrendering or are *hors de combat*—and even posing a risk to friendly forces, it is unsurprising that normative debate on the use of AI

 DOI: 10.1201/9781003410379-15

in military contexts has been focused on whether lethal Autonomous Weapon Systems (AWS) could be, or are, morally permissible and legally compliant.[1]

While these debates are right and proper, it is noteworthy that the policy debate has largely moved from whether there should be a ban on AWS, to how they should be regulated, both in terms of the principles applying to their regulation and the relevant institutional structures (e.g., Trager & Luca, 2022). Debates around the moral permissibility and legal compliance of lethal AWS suffer from a further deficit, however. Technologically developed militaries will find important uses for AI in far more varied contexts, which may pose equally severe ethical challenges, than solely for lethal AWS. While 'killer robots' are attention-grabbing, that attention has occluded proper scrutiny of other use-cases, and a serious concern that militaries should be responsible in their use of AI must address the full gamut of likely use-cases. An outline goal of this chapter is, therefore, to contribute to remedying this neglect, by considering a vital use-case—one that exemplifies the far-reaching significance of AI and which poses, I will argue, a unique challenge to the moral principles we should adopt for evaluating the proper use of AI in the military.

More specifically, the use-case I focus on is the use of AI in enabling military decision-making. In military headquarters, 'command and control' is exercised over subordinate units, directing their activities. AI is likely to play a role throughout the decision-making processes by which command and control are exercised, generating recommended courses of action, which commanders will determine whether to reject, modify, or adopt. At first glance, the use of AI in military decision-making is not as controversial as its use in 'killer robots'. The AI would not be hunting anyone, giving rise to Terminator-style fears. Moreover, humans would be not just 'on' the loop, but 'in' it, and to all intents and purposes essentially so. (As I shall use the terms, humans are 'in' the loop when humans are actively engaged in both the target acquisition and engagement processes of a combined socio-technical system, with the system requiring explicit human authorization before engaging a target, and humans are 'on' the loop when a human monitors the operation of a system, which may itself be autonomous or semi-autonomous, with the human able to intervene to stop or change the action that the system would otherwise take. The US Department of Defense terms the latter 'operator-supervised' systems (2023, pp. 22–23). I use the short-hand designations of 'off', 'on', and 'in' to refer to systems with the properties as stated here.) ChatGPT may help staff write the voluminous operational orders, but no commander will abdicate to a machine their prerogative to issue orders, not least on pain of redundancy.

While this lack of controversy is indubitable, I will argue that this belies the moral reality. A large part of the advantage that AI is likely to realize for militaries on the battlefield depends on its capacity for what I call 'unpredictable brilliance', and it is the central problem of this chapter. To illustrate this problem, I start with a lesson from DeepMind's AlphaGo programme and apply it to the military. Then I turn to evaluate its significance. I identify the risks that AI-enabled decision-support systems may pose to militaries, preparatory to showing how these systems will redistribute responsibility within militaries. Much discussion of lethal AWS has already focused on the challenge of allocating responsibility; I show how the problem of unpredictable brilliance deepens this challenge, so that human-in-the-loop systems may require the allocation of responsibility to diverse responsibility-holders, just as much as human-off-the-loop systems

do, and that the phenomenon of 'blameworthiness gaps' arises for them too. I close by considering an objection to my account, deriving from bioethics discussions of the role of AI-enabled diagnosis and treatment.

ALPHAGO'S MOVE 37

In 1997, IBM's Deep Blue became the first computer programme to beat a chess grandmaster, Gary Kasparov, and with subsequent improvements in computing power, computers now standardly beat human opponents (Kasparov, 2010). A harder challenge was therefore sought, as part of the long development of AI, and it was found in the game Go. While both are strategy games, the rules of Go are simpler than those of chess. One player has a bowl of black stones, and another of white, and they take it in turns to place them on a board. Stones of one color which are surrounded by those of the other are removed from the board, with the goal being for a player to surround a larger area, in total, than their opponent.

Although simpler than chess in respect of its rules, Go is nonetheless considerably more challenging in terms of what is required to analyze a position computationally and recommend a move. The complexity of a game like Go and Chess, where all the possible moves can be identified, can be defined in two terms, namely those of 'breadth' and 'depth'. From any single game position, a player must choose one move from a limited set of options, with the set of options its breadth. And, from a given position, a game will on average last a certain number of moves; this is its depth. A game's breadth and depth from a given position then expresses the number of branches down which the game could develop. On average, a typical chess game has a breadth of 35 moves and a depth of 80, meaning the number of possible moves that a computer might calculate, when evaluating which move will maximize the likelihood of victory, is vast, at 35^{80} (or 10^{123}). This is a huge number, exceeding the number of atoms in the universe (about 10^{80}), but still much, much smaller than Go. With Go played on a 19 by 19 board, and having a breadth of 250 and depth of 150, there are about 250^{150}, or 10^{360} moves. 'This is a number beyond imagination and renders any thought of exhaustively evaluating all possible moves utterly and completely unrealistic' (Koch, 2016; this is also the source for this quantitative comparison).

The problem of analyzing Go was, nonetheless, solved by the programme AlphaGo, the product of the company DeepMind. Using a combination of neural network algorithms and Monte Carlo tree search techniques (Silver et al., 2016), and with AlphaGo playing against itself repeatedly, the programme was able by 2015 to beat a professional Go player who was also the reigning European Go champion, Fan Hui. A series of five, highly publicized matches was subsequently set up in 2016, with a $1 million USD prize at stake, in which AlphaGo would compete against Lee Sedol, widely recognized as one of the best Go players in the world at the time. With AlphaGo having won the opening match, arguably the pivotal moment in the series occurred in the second game, in move 37, in which AlphaGo plays a 'fifth line shoulder hit'. The commentators' immediate reactions at the time were variously of shock and surprise. (The following reactions are

cited from Kohs, 2017.) "Oh, totally unthinkable move." "That's a… That's a very surprising move. Coming on top of a fourth line stone is really unusual. I wasn't expecting that." Fan Hui, who was acting as a judge for the Sedol-AlphaGo matches, commented afterward, "When I see this move, for me it's just a big shock. What? Normally humans, we never play this—because it's bad. It's just bad; we don't know why; it's just bad." Another commentator, explaining the move, remarked, "It's the fifth line. Normally you don't make a shoulder hit on the fifth line." In short, among those watching at the time, who had the expertise required for interpreting the game to a general audience and responsibility for doing so, there was considerable confusion and indeed shock at the move.

AlphaGo's analysis of the game situation suggested that the commentators' surprise at move 37 was well-founded. According to AlphaGo's own assessment, there was a 1-in-10,000 chance that a human player would have made this move. Lee Sedol's own assessment of the move agreed that it was surprising, but added something further:

> I thought AlphaGo was based on probability calculation and that it was merely a machine. But when I saw this move, I changed my mind. Surely AlphaGo is creative. This move was really creative and beautiful. This move made me think about Go in a new light. What does creativity mean in Go? It was a really meaningful move.

Although the game was perceived by some as remaining balanced between the two players for a significant period after, subsequent commentary suggested that move 37 was the crucial move that ultimately won the second game for AlphaGo. Fan Hui, again, remarked, "This move was very special, because with this move, all the stones played before worked together. It was connected. It looked like a network, linked everywhere. It was very special." AlphaGo went on to win the series by four games to one. (The prize money was donated to charity by the DeepMind team). AlphaGo Master, a subsequent iteration of AlphaGo, went on to beat the acknowledged world No. 1, Ke Jie, in 2017. Also in 2017, DeepMind reported the development of AlphaGo Zero, a further iteration of the programme, which beat the original AlphaGo 100-0 (Silver et al., 2017).

How is AlphaGo able to make such surprising decisions that its moves are described as 'creative' and 'beautiful'? Two points are, I think, noteworthy. First, I noted above that AlphaGo was able to play itself repeatedly, and thereby learn what moves are likely to increase or decrease the chance of winning. But this understates the power of machine learning techniques. For AlphaGo, this consisted of playing against itself 'thousands of times', using reinforcement learning to build on the heuristics identified in a training data-set drawn from 30 million human games (Scharre, 2018, p. 125; DeepMind, 2023). AlphaGo Zero eschewed the training data and simply played itself, 4.9 million times (Silver et al., 2017, p. 355). In both cases, AlphaGo was able to experiment, at a speed and therefore on a scale vastly beyond that accessible to the individual human player, with different permutations of moves, discarding those which proved not to conduce to victory, but preserving the lessons from those which were—with some of these being types of move that humans had discounted as unwise.

Second, while AlphaGo tries to maximize its probability of winning, it does not care about the margin by which it wins. Go permits degrees of victory, in which one player

may win by more or less, according to how much territory each player's stones control at the end of the game. Faced with the choice between an 89% chance of a substantial victory, or a 90% chance of marginal victory, AlphaGo will choose the higher-likelihood-but-marginal one. This 'willingness' to win by a small margin on AlphaGo's part is likely to contribute to its capacity to make seemingly surprising moves, with humans' preference to win by a substantial margin partly based on psychological preferences (the appeal of crushing one's opponent), and partly because the predicted degree of victory is a useful heuristic for estimating the likelihood of victory. But heuristics often fail, and AlphaGo seems able to identify those occasions when it is not reliable, and so make surprising recommendations. Of course, optimizing AlphaGo to maximize the odds of victory, rather than some weighted goal that takes account of both the odds of victory and the degree of victory, is a design decision. In the context of the game Go, the design decision is not an especially significant one, while in an unbounded, real-life context, the decision about what a machine learning algorithm should optimize on is likely to be highly consequential.

In an initial summary, then, we can describe AlphaGo's move 37 as possessing the property of 'unpredictable brilliance'. It was brilliant because it proved to be decisive, in the context of the game in which it was played, unbalancing the programme's opponent, and setting AlphaGo on a path to victory. It was unpredictable not only because the odds of a human making that same move were astonishingly slim—although they were—but also because even those watching at the time did not realize that the move would have the decisive effects on the game that it proved to and which, from AlphaGo's perspective, seeking to optimize the chances of victory, were welcome. Quite simply, it was a move that those human observers would not have chosen to have played, likely even if it had been recommended to them by an AI.

The phenomenon of unpredictable brilliance is observed in other forms of AI. While I have drawn the above portrayal of AlphaGo's move 37 from a single source, the *AlphaGo* documentary (Kohs, 2017), Paul Scharre documents other instances of the same phenomenon, to substantiate what he describes as AI's 'alien' and 'inhuman' form of cognition (2023a). The AI programme Libratus routinely beats humans in poker, and employs a different strategy, using betting tactics 'like limping and donk betting that are generally considered poor tactics, but it is able to execute them effectively because of a more fine-grained understanding of the game's probabilities' (Scharre, 2023a, p. 266). The same is true of AlphaZero's chess style, where it will even sacrifice a queen for positional advantage that pays off over the game.

I have suggested that AI's 'alien' cognition is likely to stem, in part, from the vastly greater dataset of possible gameplays than it can access which an individual human cannot, and from the fact that it may be programmed to optimize on a single outcome. But this is unlikely to be the major reason for the 'alien' nature of its cognitions, with other sources likely to be contributory as well. The efficiency gains that an AI can realize may be of such a degree that it makes decisions that feel qualitatively distinct. For instance, the company OpenAI has a programme, OpenAI Five, able to control agents in the *Dota 2* multiplayer online battle arena (MOBA). MOBA games are more complex again than games like Go, not only with players having vastly greater numbers of options open to them at any point, but also with vastly more players, playing for longer, so yielding much larger breadth and depth. In addition, information is asymmetric, with

some information hidden from other players (unlike Go or chess). The teams of virtual agents controlled by OpenAI Five are—by now, perhaps unsurprisingly—able to beat teams of professional humans, being able better to coordinate their attacks, and to do so with more speed and precision, so overwhelming opposing teams. The bots are reported to play 'with unusual aggressiveness' relative to human players (Scharre, 2023a, p. 269). In an explicitly defense-related context, a series of simulated aerial dogfights between AI-controlled and human-controlled fighter jets resulted in a 5-0 win for the AI. The AI demonstrated a strong preference for 'forward-quarter gunshots', in which two aircraft fly directly toward each other. Human pilots invariably seek a 'rear-quarter' shot—i.e., from behind their opponent—largely because the alternative is incredibly dangerous to the pilot (and therefore forbidden in training), and also incredibly difficult to pull off, because in the split-second window of opportunity that exists to hit the target, the pilot's priority is invariably to avoid a crash (Scharre, 2023a, pp. 2–3). While the AI could pull off the forward-quarter shot, in the simulation at least, the human could not. (It is a further question how well performance in the simulation predicts performance in actual air-to-air combat.) The efficiency gains that an AI can realize can come from multiple sources: with no cognitive processing limitations, an AI is likely to possess greater situational awareness, a greater ability to manage its own resources, and an ability to carry out multiple tasks in parallel. It will also be risk neutral, exhibiting neither the natural risk aversion of military personnel who wish not to die, nor the risk-seeking behavior that is an indubitable feature of war, especially in elite units, which is normally functionally valuable in overcoming risk aversion, but which can lead personnel to make foolhardy decisions.

While efficiency gains, and access to vastly greater data-sets, are two sources of the counter-intuitive nature of AI cognition, it becomes truly 'alien' when it is the result of algorithms that are themselves developed by machine-learning techniques, to the point when they become, in effect, black boxes which are not inspectable by humans.

AI's 'alien' cognition is likely to be exhibited wherever it is found, and this includes one of the most significant use-cases for AI in the military, namely as an aid to decision-making. Part of the battlefield advantage that AI can offer will consist not just in individual platforms that can outcompete the equivalent human-controlled adversaries, as in a contest between an autonomous air-superiority fighter and one piloted by a human, but also in decisions about when and how to deploy which forces in such a way that one side can compel another to submit. In military headquarters, at levels spanning the tactical, operational, and strategic (i.e., from the battle group, in the land domain, or the ship, in the maritime, up to national or multi-national theatre commands), large staffs evaluate information which may reveal what the enemy is doing, in order to generate intelligence; maintain an awareness of the location and state of friendly forces; generate recommendations to commanders on possible courses of action; and once decisions are taken, then issue orders to subordinate units and track their implementation. All of this is summarized as the exercise of command and control. AI is likely to play a role throughout all elements of this decision-making process, and a transformative role in two parts in particular: namely, in generating intelligence, and in generating recommendations to commanders on possible courses of action, both for deliberate, pre-planned actions and in time-sensitive contexts. For present purposes, it is the latter which is of most interest.

Possessing far greater situational awareness, with an AI-enabled intelligence picture that is less susceptible to confirmation bias, AI-enabled decision-support systems are likely to give recommendations that, in some instances, confirm what a human commander would be likely to choose or modestly improve it, but in some instances, will offer highly surprising and counter-intuitive recommended courses of action (COAs). Just as much as individual platforms, like autonomous fighter jets, AI-enabled decision-support systems are likely to exhibit the property of unpredictable brilliance, with COAs that sometimes seem to be simply a waste of resources, and at other times also dangerous. And yet, this is likely to be precisely where the most significant advantage lies, for those forces that can field powerful AI-enabled decision-support systems, and effectively integrate their insights into their command and control. Decision-making which is both counter-intuitive—indeed, creative—and highly rational, being based on a sophisticated understanding of the probable outcomes from a range of possible actions, is likely to be able to exploit any errors an enemy has made in its own force disposition, and to unbalance the adversary. As Scharre writes, in summary, 'The militaries that will be most successful in harnessing AI's advantages will be those that effectively understand and employ its unique and often alien forms of cognition' (Scharre, 2023b; also Scharre, 2023a, p. 273).

RISKS POSED BY UNPREDICTABLE BRILLIANCE

Unpredictable brilliance is not just a property of AI cognition, but also a problem. The next section considers the problem that it poses in moral terms, but I start here with the problem that it poses for commanders. For commanders, the problem can be easily stated, at least in outline terms: to what extent should the AI-generated recommendation be trusted? In the context of AlphaGo's games against Lee Sedol, this problem did not arise, in part because the real-world stakes could effectively be discounted (the prize money was a trivial amount for a company like Alpha, Google's parent company and the owner of DeepMind, which had probably put up the money for publicity-related reasons), but largely because the game was set up so that the AI's recommendations determined what move was played. While AlphaGo was purely a piece of software and relied on a human player to place a black or white stone on the physical board, that person nonetheless had no more role other than enacting AlphaGo's decision as it was represented on a display screen. The series was an experiment, and trust played no part in it. But in the military context, the stakes would be substantial, and the human commander would not be required simply to follow an AI-enabled decision-support system's recommendation. Instead, the AI would generate recommended COAs, not instructions, and the commander would ultimately be accountable for the consequences of her decisions. She therefore faces a tricky decision: given an AI-recommended COA that has obvious risks, and for which the benefits are not easily discernible, should she follow the recommendation, or her own instincts and judgment? In effect, whom should she trust—the AI, or herself? Doing the latter means that she would be able to defend her

decision, by the lights of accepted canons of military decision-making wisdom. But the consequence is that she would forego the prospect of a decisive battlefield advantage held out by the AI-enabled support.

Trust is necessary in part because, in the military context, there is significant risk. There are two forms of risk that are especially noteworthy, both of which arise because an AI-recommended COA could be unpredictable, or surprising, due to how it would allocate military resources on the battlefield. In more concrete terms, the COA may be surprising because it recommends deploying a type of unit that is unusual in a given context—for instance, it might recommend sending a main battle tank where normally dismounted infantry would go. (What would its reason for doing so be? Perhaps the AI-generated intelligence picture indicates an extremely low likelihood of dismounted 'red' infantry in the immediate urban environment, but the combined speed and firepower of the tank will be decisive in dislodging the enemy from the fixed positions on the other side of the town, which the enemy is in the process of preparing now, and which positions seek to deny a crucial bridge to 'blue'. But the assessments and probabilities taken account of in the AI's decision-making process may be hidden from the commander.) Or, its recommendation may be surprising because it would result in a significant amount of combat power being invested in a given location, which the commander and her staff assess as being unimportant, but which the AI assesses as being key terrain. Given that military commanders' decisions, ultimately, address the allocation of military resources, which equates to combat power, across the battlefield, the property of unpredictable brilliance would be realized through surprising, counter-intuitive recommendations about where that resource should be placed. One form of risk that a commander would have to accept, in following the AI's recommendation against her own judgment, is that which the decision would pose to the overall likelihood of achieving her campaign objectives. Whether this is at the tactical, operational, or strategic levels, in a world of scarce resources, if military assets are misallocated, the relevant objectives are less likely to be achieved. This is a risk that she should be highly sensitive to.

Another form of risk is that which the decision poses to the lives and safety of blue force personnel. This is especially obvious when an AI-generated recommendation would see personnel or crewed platforms deployed in ways where they would be significantly more vulnerable than would be the case for personnel deployed in the range of decisions that would be likely for a human commander. As implicitly suggested above, main battle tanks and other armored vehicles are particularly vulnerable to dismounted infantry in 'close' country—forests, mountains, and urban environments in particular. Conversely, dismounted infantry are particularly vulnerable in 'open' countries, such as plains and deserts. Deploying one form of combat power in terrain that is widely considered more suitable for the other is likely to lead to significantly greater casualties, at least according to widely accepted military heuristics. (A significant proportion of Russia's casualties during the first year of its invasion of Ukraine in 2022 is explained by poor decisions about how to deploy armored vehicles, which failed to follow the heuristic noted above.) Similar points apply to aerial and maritime platforms, which have differing combinations of firepower, mobility, and resilience, and therefore differ in how they are usually used effectively.

Although it is plain that an AI-generated recommendation would pose high levels of risk if personnel were ordered to go into situations that standard tactics, techniques, and procedures (TTPs) indicate would be unduly dangerous for them—being evident, not least, to those personnel themselves—the point is actually a more general one. Any situation in which an AI-generated recommendation is significantly discrepant from the range of likely decisions for a human commander involves a redistribution of risk, in which some people have a higher level of risk imposed on them. Where the baseline for comparison is the likely range of decisions that a human commander would make, and the AI's recommended COA is outside of this, the AI's COA will result in a set of some blue force personnel carrying less risk—perhaps because they are now part of a larger formation, or there is a larger reserve held back, and so on—but almost inevitably at the cost of an increase of risk for a set of some other blue force personnel. Even though the overall amount of risk is not fixed (because some COAs pose a greater level of risk overall, while others are less risky), nonetheless, one person's gain, in terms of risk reduction, is likely to be due to some other person's loss, in terms of increased risk. While that risk redistribution may be justified, perhaps in terms of its impact on the overall risk to blue force personnel, or its increase in the odds of mission success, nonetheless it will have occurred, even though that risk redistribution may also be invisible to those personnel who are affected. By hypothesis, when an AI's recommended COA is unpredictable or surprising, that COA is assessed by humans as riskier, in terms of its probabilities of either or both of mission success or blue force casualties, than the likely range of human-recommended COAs. (If it was not so assessed, it would be within the likely range of human COAs). What is unknown *ex ante*, and indeed, is unknowable, is whether the AI's recommended COA imposes, objectively, less risk to blue force personnel overall, and has a higher probability of mission success, than the likely range of COAs that a human would adopt, or whether the contrary is the case.[2]

I have claimed that trust would be necessary, when a military commander considers whether to adopt a COA recommended by an AI-enabled decision-support system, in part because there is significant risk. It may be objected, however, that this overstates the case. Although risk is inevitably involved in any COA adopted on the battlefield, the commander's trust of an AI-enabled decision-support system would be greatly minimized because she would not be interacting with the system from the beginning, but rather would have trained with it on multiple occasions. She would have a track record of its performance, enabling her to evaluate how much it should be trusted, under what conditions, and so on. This is of course correct; it would be highly unprofessional for a military to deploy an untested piece of equipment on an operation. Nonetheless, while this may mitigate the need for trust, it does not eliminate it, and the fundamental choice for a commander of trusting her own instincts, versus the recommendation(s) of an AI, is likely to remain. However much training has been conducted, a decision-support system is different from more conventional kinds of military equipment, such as platforms or missiles, in that its utility essentially turns on how accurately it represents and predicts an adversary's actions. Operations are qualitatively distinct from training because, in training, you actually fight against your actual enemy. Only a 'two-way range' permits that. As it is sometimes said, on operations, the enemy has a vote. Nor is this need for trust restricted to the start of operations on a given campaign. Rather, the need for trust will be recurring. War is one of the fastest drivers of human innovation, so as the

red force adapts its TTPs and equipment to respond to blue force's TTPs and equipment, the COAs that will successfully unbalance and exploit red force weaknesses will be correspondingly new and, in some instances, counter-intuitive.

UNPREDICTABLE BRILLIANCE ERODES THE MORAL SIGNIFICANCE OF THE DISTINCTION BETWEEN HUMANS *OFF* AND *IN* THE LOOP

As noted earlier, much and perhaps most of the ethical debate about the use of AI in military contexts has revolved around lethal AWS, or 'killer robots', in which humans are out of the loop. The intuitive thought is that fully autonomous systems pose unique challenges, as contrasted with systems in which humans are on or in the loop. A central reason why one may be concerned about lethal AWS is that they seem to raise the prospect of killings in war, especially of those not liable to be killed, such as non-combatants or surrendering soldiers, for which no-one is clearly responsible. If humans are in the loop or on it, they can be held responsible; if humans are off the loop, it is less clear, and some have argued that in principle no-one is (Matthias, 2004; Sparrow, 2007).

While the distinction between autonomous systems and those which lack full autonomy is clear enough, its moral import is less significant than it seems. One of the tactical advantages of AI-enabled weapons systems is the speed at which they can operate. It is very plausible that systems might be developed in which humans remain in or on the loop, thus retaining the seeming moral advantages of having an identifiable person who can be held responsible for each targeting decision, while putting that person in a situation in which, given that the tactical situation is evolving at such speed, she has in effect no choice but to authorize the decision recommended made by the AI. A system that might satisfy this description is one which is designed to provide point defense against swarms of kamikaze drones, which the Phalanx Close-In Weapon System (CIWS) provides a model for. Mounted on ships to defend against incoming missiles by putting up a 'wall of lead', a human operator either approves fire recommended by the Phalanx, or the system can operate 'weapons free', engaging targets from 1.5 to 5.5 km away, with the operator monitoring the relevant conditions. Suppose the range was much tighter and the number of threats much greater, so that a future version of such a system was capable of and needed to engage 1,000 targets within five seconds from a range of 0 to 500 m, while a human operator was monitoring how the system operated. We may even suppose, in this counterfactual scenario, that the operator's express approval was continuously required—she must hold down a button for the system to strike incoming targets. The operator would be in the loop. But it is highly dubious that she would be responsible for those strikes, in the way required to hold her individually liable for any mistakes. *De facto* and *de jure* responsibility would diverge.

Putting soldiers, sailors, or aircrew in this kind of situation could be useful for militaries, in terms of how legal liabilities are distributed, by allowing the military to hold an individual liable for illegal or reputationally damaging targeting decisions, and so

avoiding corporate responsibility. But it would be exploitative, as the soldier would lack the time or situational awareness to make an informed decision. Decision-making which is necessarily done at extremely high speed, then, and in reliance on system-generated recommendations, erodes the moral import of the distinction between humans off and in the loop. While the distinction provides a useful heuristic to distinguish between situations in which a human is individually responsible for specific targeting decisions and those in which someone is not, it is not determinative. There are cases where humans in the loop can nonetheless lack individual responsibility.

The phenomenon of AI's unpredictable brilliance, I contend, has the same erosive effect. The speed at which a decision may necessarily have to be made is one factor that can have the effect of removing responsibility from a human 'decision-maker', the scare quotes indicating that it is questionable to what extent the human is, indeed, the decision-maker. She retains the freedom, overall, to substitute her own judgment for that of the AI, but the consequences of doing so are likely to be severe. In the case of AI support to commanders in tactical, operational, and strategic headquarters, the same erosion of responsibility occurs. Faced with an unpredictable or surprising COA recommended by an AI, the military commander retains a decision as to whether to follow the AI's recommendation or to follow her own. All she has to guide her in that decision is whatever information she possesses about the reliability, in general, of this AI's recommended COAs—there is no more granular level at which she can assess this proposal because, by hypothesis, the proposal runs counter to the accepted principles of military decision-making. Further, determining the overall reliability of an AI is not, primarily, her responsibility, but the responsibility of the force development and procurement programmes that brought the system into service and approved it for operational deployment. So long as she uses it in a way that is compatible with the directions on appropriate use, it is by no means clear to what extent she can be held individually responsible for a decision, the consequences of which turn out to be bad.

What follows from this? It does not follow, at least not straightforwardly, that there is an in-principle, moral objection to the use of such AI-enabled decision-support systems. In other work, I have argued that, subject to some constraints being satisfied, lethal AWS could be permissibly deployed (Simpson & Müller, 2016). The crux of the issue is not whether someone identifiable is responsible for each killing by a lethal AWS. There are multiple roles the occupants of which may be responsible for such killings, including those who design and maintain lethal AWS, those commanders who deploy them on a given occasion, and most especially, those who authorize the use of a given system and regulate its use. And it is always the case that someone—either an individual or a corporate body—is responsible. But responsibility does not imply blameworthiness. It is wholly possible that there could be some killings by lethal AWS, which in technological terms would be classified as malfunctions and in human terms as tragedies, in which someone is killed who is not liable to be killed, like a non-combatant, and yet for which no-one is blameworthy. This is because the use of lethal AWS may satisfy the demands of 'wide proportionality', in which harms to those who are innocent are weighed against an act's expected good effects (McMahan, 2020, p. 15). Whether the use of a system satisfies the demands of wide proportionality may occur on occasions in which a non-combatant's predicted death is a proportional cost for a given strike on

a valuable military target. Although that point is not controversial, it is also noteworthy that a system may satisfy those demands and yet impose the risk of harm to the innocent in a wider range of situations, such as a non-combatant being mistakenly identified by an AWS as a combatant. For lethal AWS to be used proportionately, however, it is not sufficient simply that the absolute level of risk of humans should be reduced. Rather, we should also be sensitive to how the use of lethal AWS would redistribute risk. In particular, it is not morally acceptable for AWS to reduce risk to blue force personnel but at the cost of increasing it to people who are not liable to be killed. It is a fundamental constraint on the just use of lethal AWS that the risk they pose to non-combatants should be lower than would be posed by a military not equipped with lethal AWS and, further, that the risk should be as low as is technologically feasible (Simpson & Müller, 2016).

This account of how responsibility and blame are allocated applies not only to lethal AWS, but also the use-case of concern here, of decision-support systems. Suppose that a surprising, counter-intuitive COA is recommended by an AI. Even though a human would be in the loop, it does not follow that that commander carries individual responsibility and blame for adopting that recommended COA. The surprising and counter-intuitive features of AI cognition erode her responsibility. So long as she is integrating the system into her command and control in accordance with the principles and guidance she has received, the commander is absolved from blame for the consequences. Those on whom responsibility principally rests, and on whom blame is most likely to fall, if it does, are those individuals or that corporate body which has authorized the use of the AI-enabled system and regulates its use. And it is possible that no-one should be blameworthy, even if the system results in harmful risks eventuating, so long as the demands of wide proportionality are met. Positively, then, what follows from AI's property of unpredictable brilliance is that there will be a more widespread distribution of responsibility, and liability to blame, within the military, away from commanders, and toward the institutions (both individual office-holders and as corporate bodies) that authorize and regulate the systems by which wars are likely to be fought in future, included in which are AI-enabled decision-support systems. The widespread distribution of responsibility is already a feature of the military, with established procedures and institutions responsible for testing, evaluating, validating, and verifying new equipment and TTPs. My contention is that, with the introduction of AI-enabled decision-support systems, the widespread nature of this distribution will become more accentuated still. At present, commanders retain a core responsibility for 'J3' decisions when deployed—those decisions that address directly how operations should be conducted. Yet even this is likely to be restricted. As a parallel, and in a use-case I consider further below, note the likely implications if AI-enabled diagnoses of illness and recommendations for treatment achieve better health outcomes than medical doctors. Doctors' responsibility then would be to assure themselves that the AI-enabled decision-support system is being used in a context in which it is appropriate, but once done, the individual's responsibility would be simply to adopt the AI-generated recommendation. Similarly, once a military commander has confirmed that the context in which she is operating is suitable for the decision-support system, little will remain for her to do other than accept the AI's recommendation. Her command responsibility will have been redistributed, to the procedures and institutions responsible for developing and maintaining the AI.

To make the point plain that, even with a human in the loop, a commander can nonetheless be absolved of blame, I have started with scenarios in which an AI makes a surprising, unpredictable recommendation. This is an expository device, however, to support a more general and perhaps less intuitive conclusion. Operationally deploying an AI-enabled decision-support system would have the effect of absolving the human commander from blame not only in situations in which the AI's recommended COA was outside the range recommended by humans, but also in situations in which it was within. The effect of authorizing a system for operational deployment, with responsibility for its reliability resting with those performing the authorization, is to remove blameworthiness as such for its decisions from commanders. Not only would the commander be absolved of blame, but as a corollary, she would also be absolved from praise.

The likely effects of this on how militaries function, in sociological and organizational terms, should not be underestimated. For centuries militaries have valorized and praised commanders who show an instinctive understanding of the battlefield, often disdaining personal risk, and have led their soldiers to great victories—in Europe, think of Alexander the Great, Napoleon, Nelson, Rommel, and Patton. As and when AI becomes not just embedded in the decision-making process, but the central driver of it, we will enter a post-heroic era of warfare, where leadership may be exercised at the lowest tactical levels, at which soldiers must still face risk and overcome their fear, but it will not be exercised at the operational level. The social dislocation will be profound.

AI DECISION-SUPPORT IN HEALTHCARE: AN OBJECTION

I conclude with an objection. Despite my foregoing account of how responsibility and blame are allocated for AI-enabled systems, could there nonetheless be an in-principle, moral objection to their use? Consider the related context of healthcare, in which AI-enabled decision-support systems can advise a doctor on possible diagnoses and treatments for a patient—a domain where, similarly, lives depend on the quality of decision-making. Examples include IBM's 'Watson for Oncology' and Aidoc Medical's triage and notification systems. A number of writers in bioethics have pointed to other moral risks associated with AI-enabled medical decision-support systems, in addition to the redistribution of responsibility that they impose (Grote & Berens, 2020, p. 209), even though such systems may achieve better results than a human doctor. It is claimed that they undermine trust in the doctor–patient relationship (Hatherley, 2020) and may conflict with core ideals of patient-centered medicine, undermining patients' autonomy, resulting in treatment no longer being the outcome of shared deliberation between both parties (Bjerring & Busch, 2021; Lorenzini et al., forthcoming; McDougall, 2019). Although these authors have differing accounts of the significance of these moral risks, their shared point is that while effective treatment is one valued feature of healthcare, it is not the only one and that these valued goals may trade-off against each other.

These moral risks of AI-enabled medical decision-support undoubtedly exist. In the context of healthcare, it is also plausible that these risks could be the basis for principled limits on the use of AI-enabled decision-support systems. But this derives from a core feature of medical practice, namely the consent of those exposed to risk, that does not have the same significance in the military situation. In healthcare, patients face some harm, in the form of injury or illness, and they undergo treatment with its attendant risks to improve their welfare. If a patient ignores an AI's diagnosis and recommended treatment because they wish to know the basis of that assessment, so accepting a risk of worse outcomes for the sake of preserving their autonomy, that is their trade-off to make. But in the military context, while consent is exercised when one joins up (at least in professional militaries), it is not relevant when deployed. Subordinates are under the authority of their superiors, duty-bound to follow orders, and part of what they have consented to when joining up is a liability to assume high levels of personal risk in the course of carrying out operations. In turn, superiors have a duty of care to ensure that the personal risk their subordinates are exposed to is only that which is necessary for the military, collectively, to achieve its campaign objectives. In the military context, then, commanders are not free to trade-off better campaign outcomes against 'softer' values, such as ensuring that the basis of a decision can be explained. Commanders' core goal is to achieve their campaign objective, and in doing so they are to minimize the risk to their personnel, and civilians. Insofar as an AI's recommended COA will enable them to do so better, and this is known within their military, commanders have a duty to comply with that recommendation.[3]

NOTES

1 While it is widely believed that weapons systems designed to kill and which are capable of operating autonomously exist, it is not publicly known whether AWS have in fact been used to kill a human autonomously. Robert Trager and Laura Luca state that "at least Israel, Russia, South Korea, and Turkey have reportedly deployed weapons with autonomous capabilities" (Trager & Luca, 2022). A UN report stated, in passing, that the Turkish STM Kargu-2 was used in Libya while 'programmed to attack targets without data connectivity between the operator and the munition: in effect, a true "fire, forget and find" capability', but this capability was denied by the manufacturer (United Nations Security Council, 2021, §63; Bajak & Arhirova, 2023). It is likely that the first uses of lethal AWS will not be publicly known about at the time, and likely for some time after, as there is no easy way to determine after a strike whether, for instance, a loitering munition, such as the Switchblade 'kamikaze' drone operated by the US or the Zala Lancet operated by Russia, was used with a human in the loop, or off it.

2 To make things even more complex, not only is the objective risk posed by an AI's COA unknowable *ex ante*, compared to the range of likely human

COAs, it is also unknowable *ex post*. That COA 1 was adopted means that COA 2 was not, and the alternative reality in which COA 2 was adopted is unavailable.

3 I am grateful to Jan Maarten Schraagen for comments and criticism.

REFERENCES

Bajak, F., & Arhirova, H. (2023, January 3). Drone advances in Ukraine could bring dawn of killer robots. *Los Angeles Times*. https://www.latimes.com/world-nation/story/2023-01-03/drone-advances-in-ukraine-dawn-of-killer-robots

Bjerring, J. C., & Busch, J. (2021). Artificial intelligence and patient-centred decision-making. *Philosophy & Technology, 34*, 349–371. https://link.springer.com/article/10.1007/s13347-019-00391-6

DeepMind. (2023, June 26). *AlphaGo.* deepmind. https://www.deepmind.com/research/highlighted-research/alphago

Grote, T., & Berens, P. (2020). On the ethics of algorithmic decision-making in healthcare. *Journal of Medical Ethics, 46*, 205–211. https://jme.bmj.com/content/46/3/205.abstract

Hatherley, J. J. (2020). Limits of trust in medical AI. *Journal of Medical Ethics, 46*, 478–481. https://jme.bmj.com/content/46/7/478

Kasparov, G. (2010, February 11). The chess master and the computer. *The New York Review.* https://www.nybooks.com/articles/2010/02/11/the-chess-master-and-the-computer/

Koch, C. (2016, March 19). *How the computer beat the Go master.* Scientific American. https://www.scientificamerican.com/article/how-the-computer-beat-the-go-master/

Kohs, G. (Director). (2017). AlphaGo [Film]. Moxie Pictures; Reel as Dirt.

Lorenzini, G., Ossa, L. A., Shaw, D. M., & Elger, B. S. (Forthcoming). Artificial intelligence and the doctor-patient relationship expanding the paradigm of shared decision-making. *Bioethics*. https://onlinelibrary.wiley.com/doi/full/10.1111/bioe.13158

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology, 6*(3), 175–183. https://link.springer.com/article/10.1007/s10676-004-3422-1

McDougall, R. J. (2019). Computer knows best? The need for value-flexibility in medical AI. *Journal of Medical Ethics, 45*, 156–160. https://jme.bmj.com/content/45/3/156.abstract

McMahan, J. (2020). Necessity and proportionality in morality and law. In C. Kreß & R. Lawless (Eds.), *Necessity and proportionality in international peace and security law* (pp. 3–38). Oxford University Press. https://academic.oup.com/book/33456/chapter/287728709?login=true

Scharre, P. (2018). *Army of none: Autonomous weapons and the future of war.* New York: W. W. Norton

Scharre, P. (2023a). *Four battlegrounds: Power in the age of artificial intelligence*. New York: W. W. Norton

Scharre, P. (2023b, April 10). *AI's inhuman advantage*. War on the Rocks. https://warontherocks.com/2023/04/ais-inhuman-advantage/

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Driessche, G. van den, Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D, Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., & Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature, 529*, 484–489. https://doi.org/10.1038/nature16961

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., Driessche, G. van den, Graepel, T., & Hassabis, D. (2017). Mastering the game of Go without human knowledge. *Nature, 550*, 354–359. https://doi.org/10.1038/nature24270

Simpson, T. W., & Müller, V. C. (2016). Just war and robots' killings. *Philosophical Quarterly, 66*(263), 302–322. https://doi.org/10.1093/pq/pqv075

Sparrow, R. (2007). Killer robots. *Journal of Applied Philosophy, 24*(1), 62–77. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1468-5930.2007.00346.x

Trager, R., & Luca, L. (2022, May 11). Killer robots are here - and we need to regulate them. *Foreign Policy.* https://foreignpolicy.com/2022/05/11/killer-robots-lethal-autonomous-weapons-systems-ukraine-libya-regulation/

United Nations Security Council. (2021). *Final report of the panel of experts on Libya established pursuant to Security Council resolution 1973 (2011).* https://documents-dds-ny.un.org/doc/UNDOC/GEN/N21/037/72/PDF/N2103772.pdf?OpenElement

US Department of Defense. 2023. *DoD Directive 3000.09: Autonomy in weapon systems*. https://www.esd.whs.mil/portals/54/documents/dd/issuances/dodd/300009p.pdf