

Design for responsibility: safeguarding moral perception via a partnership architecture

Tjerk de Greef · Alex Leveringhaus

Received: 13 December 2013 / Accepted: 16 February 2015 / Published online: 15 March 2015
© The Author(s) 2015. This article is published with open access at Springerlink.com

Abstract Advanced warfare technologies (AWT) create unprecedented capabilities to control the delivery of military force up to the point, some argue, that we are loosing humanity. But dependence on them generates difficult moral challenges impacting the decision-making process, which are only beginning to be addressed. In order to arrive at an informed opinion about the impact of AWT on decision-making, we need to know more about what AWTs are and how they operate. We provide a short overview of the different types of AWTs and discuss the key principles that underlie Humanitarian Law. We also discuss the impact of physical distance and increased levels of autonomy on AWT and discuss the challenges posed to moral perception. Before such systems can be deployed, we need to rest assured that their usage enhances, rather than undermines, human decision-making capacities. There are important choices to be made, and sound design is ‘design for responsibility’. As a solution, we therefore propose the partnership architecture that embeds concurrent views of the world and working agreements, ensuring that operators use appropriate information in the decision-making process.

Keywords Sensemaking · Drones · Unmanned systems · Moral perception · Responsibility · Partnerships · Working agreements · Human–machine teamwork

T. de Greef (✉)
Interactive Intelligence Group, Delft University of Technology,
Mekelweg 4, 2628 CD Delft, The Netherlands
e-mail: t.e.degreef@tudelft.nl

A. Leveringhaus
Oxford Institute for Ethics Law and Armed Conflict & Oxford
Martin School, University of Oxford, Oxford, UK

1 Introduction

Today’s use of drones is no longer limited to surveillance and reconnaissance tasks only. Drones are also deployed for the uncontroversial practise of targeted killings. Typically, a drone is launched in the vicinity of an area of interest but controlled from a remote control cabinet, often thousands of miles away. In addition, military organizations consider equipping drones with high levels of autonomy in order to reduce the degree and amount of interactions between them and their operators, thereby lowering the cost of ownership.

One issue within the academic and public discourse over military actions that has not yet received sufficient attention concerns the impact of advanced warfare technologies, like drones and other robots, on the human decision-making process. There is a growing concern that ethical values are violated as a result of high levels of autonomy and remote control. Scholars from different disciplines expressed concerns regarding implication on the human decision-making process due to the distance from the battlefield and increased levels of autonomy, potentially leading to the abdication of responsibility. The impact on the decision-making process also raises concerns with regard to the principles of distinction, proportionality and necessity, all critical elements of International Humanitarian Law.

Some scholars provide arguments in favour of such advanced technologies in warfare (Altman and Wellman 2009; Strawser 2013). Firstly, for reasons of proportionality, they argue, the use is desirable *if other means are equal*. Using unmanned aerial vehicles to carry out, say, targeted killings leads to less damage and destruction than a large-scale military operation with boots on the ground. Secondly, states are under increasing pressure to minimize

casualties amongst their own service personnel, especially during operations that are not directly classifiable as self-defence (Shaw 2005). Indeed, one of the morally and politically attractive features of advanced warfare technologies is their ability to protect the lives of service personnel (Strawser 2010).

Needless to say, there have been critical voices. Ignatieff seems sceptical about the prospects of ‘virtual war’, while Robert Sparrow argues that (some types of) advanced warfare technologies must not be deployed because they undermine a commitment to moral and legal responsibility in the armed forces (Sparrow 2007). On top of the responsibility gap, Noel Sharkey is worried about so-called *killer robots* as misuse might arise... ‘by extending the range of legally questionable, targeted killings by security and intelligence forces’ (Sharkey 2010, p. 369). Furthermore, Royakkers discusses emergence of the cubicle warrior that gets morally disengaged (Royakkers and van Est 2010). These critical voices are typically used as reasons not to deploy these advanced warfare technologies.

However, we stress the importance of this highly relevant debate, and we appreciate the arguments provided by the critics and proponents. But instead of rejecting advanced warfare technologies, we use those arguments to design better systems. This requires an informed opinion about the impact of advanced warfare technologies on the human decision-making process. Following value sensitive design rationales, it is important to (re)design advanced warfare technologies for ethical values like responsibilities, proportionality, necessity and discrimination. This will lead up to advanced warfare technologies that behave more like a team player or partner. Research in this area is in its infancy, but is likely to become more prominent. We begin by giving an overview of the different types of advanced warfare technologies and discuss some of the challenges they pose to moral perception. The issue of moral perception, in fact, is crucial for a commitment to responsibility within the armed forces. We continue by making design recommendations. There are important choices to be made, and sound design is always ‘design for responsibility’—or so we shall argue.

2 Advanced warfare technologies

Since World War II, advanced warfare technologies are an important aspect in warfare, largely due to the technology readiness of advanced information and sensor technologies. It goes without saying that these advanced warfare technologies, including phased array radars, missile guidance systems and military robots, are considered an important asset in military missions. Over time, it has become clear

that such advanced warfare technologies have changed, and will continue to change, the character of war. Commenting on NATO’s intervention in Kosovo in 1999, the academic and politician Michael Ignatieff speaks of a ‘virtual war’, where NATO forces, supported by advanced warfare technologies and more traditional air power, did the fighting, but only ‘*Serbs and Kosovars did the dying*’ (Ignatieff 2000). NATO’s service personnel, Ignatieff shows, was removed from the actual combat zones, but carried out military missions with the help of technology. More recently, French, British and American warships shelled targets in Libya to assist rebel fighters in the overthrow of the late Colonel Gaddafi. Like Kosovo, intervening forces relied heavily on airpower and the latest advanced combat technologies. Today’s advanced warfare technologies increasingly render ‘*boots on the ground*’ unnecessary. This section briefly discusses a number of contemporary technologies that are already used by military organizations or that have the potential of being used in the near future (within 5–10 years).

One category of advanced warfare technologies falls with those that aid military personnel with the *dull, dangerous and dirty* work, also known as the 3D’s. The purpose of these robots is to have robots execute those tasks that are just too boring for humans, those that significantly reduce the risk associated with the task, or those that take over the dirty bits of work, including carrying a heavy load thereby reducing fatigue. The high-level goal is to have the human operator remain effective while not experiencing cognitive underload, extreme boredom or extreme levels of stress resulting from activities that directly endanger their life. Examples in this category are robots used for mine or bomb disposal purposes, deep sea missions, urban combat reconnaissance (e.g. the Dragon Runner), load carrying robots (e.g. the Alpha Dog) and powered exoskeletons (e.g. HULC).

The second category of robots concern tele-operated or unmanned vehicles. The most controversial example in this category is the unmanned aerial vehicle, which is also known as a drone given the constant, monotone and humming sound arising from the propeller engines. Instead of teleoperated or unmanned vehicles, military organizations prefer to refer to such unmanned systems as remotely piloted vehicles to emphasize that these vehicles are actually operated by specialized pilots, although not from the cockpit but from a different location, which is often thousands of miles away. Typically, a drone is launched in the vicinity of an area of interest but controlled from a remote control cabinet using a satellite relay system. In today’s military operations, the most frequently deployed drone is the Predator or the Reaper drone, both capable of carrying a payload. Although there are many different types of drones, it falls outside the scope of this paper to

list all the different types. Although the deployment of unmanned aerial vehicles is controversial for the practise of targeted killings and signature strikes, there are considered a valuable asset for surveillance and reconnaissance tasks. In addition to the group of unmanned aerial vehicles, there are also a large number of tele-operated vehicles on the ground (UGV), on the surface of the water (USV) and under water (UUV).

Many military organizations consider equipping drones with high levels of autonomy in order to reduce the degree and amount of interactions between the machine and its operators, thereby lowering the cost of ownership. For example, Rolls Royce, GE Aviation systems, QinetiQ, the UK Ministry of Defence and BEA systems are developing the Tanaris drone, named after the Celtic God of thunder. Among other ambitious characteristics, the Tanaris uses a set of stealth technologies to reduce radar reflections and emissions and has the capability to fly intercontinental missions. But, and this is important in the light of this paper, the Tanaris can act semi-autonomous. The website of BEA Systems offers a leaflet describing a future field test where the Tanaris would search an area of interest using a pre-programmed flight path to locate and identify a target (BEASystems 2011). Once identified, Tanaris seeks acknowledgement with mission control to neutralize the target. This description would fit in a category where the human would still be ‘on the loop’, where Tanaris would be flying parts of the mission without human input. On the other hand, the X47B, developed by Northrop–Grumman, has shown capabilities to take-off and land without human intervention on sailing aircraft carriers. Also, the X47B has capabilities available to refuel without human interventions.

Clearly, operational unmanned and tele-operated vehicles are, from a technology perspective, on the path of higher levels of autonomy, and it seems a matter of time, notwithstanding the effective campaign against the ‘Killer Robot’ (Human Rights Watch 2012) that led the United Nations to express a moratorium on autonomous systems (Heyns 2013a), that at least parts of the missions will be done autonomously. However, in order to have those systems be bound by the Laws of Armed Conflict (LoAC) and other moral principles, a discussion is necessary between military organization, human rights groups and LoAC representatives to link different contexts with acceptable levels of autonomy, which go beyond ‘Killer Robots’ utilized definition of autonomous systems that ‘*once activated, would select and engage targets without human intervention*’. Clearly, such a message lacks context. Context helps to differentiate between (in)appropriate use of drones in Yemen and (in)appropriate use of autonomous systems aboard a navy vessels the context of a defensive action.

The third category relates to systems that defend. The defensive systems category differentiates from the other two in that both military organizations and human rights groups accept that such systems *can* work in a full autonomous mode, and some are actually working in full autonomous mode, also when engaging. The main goal of such defensive systems is to act quickly to incoming threats thereby minimizing the risks of direct impact or fallout of debris. Acting quickly to incoming threats is not only dependent on a proper identification and classification of the incoming threat, but also on the effective ranges (i.e. weapons envelopes) of the counter-measure weapons. Typically, reaction times are short. Imagine being the chief commander aboard a modern navy frigate, alike the Dutch Air Defense and Command Frigate, the Arleigh Burke class or the Type 45 guided missile destroyers. Using its ultramodern sensor suite, several anti-ship missiles can be detected around the horizon at roughly 30 km (the horizon is dependent on the earth curvature and height of the observer). Typically, an anti-missile flies just below the speed of sound, say 300 m/s. This gives thus 100 s before impact and clearly, countermeasures need to be deployed earlier. Once an anti-ship missile is detected, a layered approach is applied starting with electronic countermeasure and decoys, followed by attempts to shoot down the anti-ship missile by short-range missiles (e.g. the Sea Sparrow or Rolling Airframe Missile). The last resorts are close-in-weapons, such as the Phalanx or the Dutch Goalkeeper system. Many navies train their officers to respond effectively to such short response times. Only when the last resort is left, being the close-in-weapons, full autonomy is accepted as those systems outperform human operators in classifying, identifying and neutralizing hostile military objects. In particular, when decisions about the delivery of force are made under pressure due to very short reaction times, having a human operator in the loop may unnecessarily delay the delivery of defensive force.

In addition to naval defensive systems, we also want to discuss the Iron Dome system, which is a defence system capable to detect rockets at 4–70 km which trajectory takes those rockets to populated Israeli territory. The Israeli Iron Dome System can itself make decisions about targeting hostile missiles, though, currently, the decision (launch of a counterattack) has to be approved by a human operator. Although the typical rocket shelled isn’t flying around the speed of sound giving the human operator significantly more time to respond the incoming threats, the use-case is that the number of incoming rockets will be higher compared to navy defensive systems that typically deal with one or two at a time.

Reduced interaction and short response time are not the only arguments in favour of higher levels of autonomy in defensive systems. Moreover, so is argued, those defensive

systems typically operate in restricted environments, such as demilitarized zones. South Korea, for example, deploys the Surveillance and Guard Robot in the demilitarized zone between South and North Korea. The robot is capable of tracking multiple moving targets using infrared and visible light cameras, and is under the control of a human operator. But the robot has the capability to identify and shoot targets automatically from over two miles (3.2 km) away. Moreover, the robot is equipped with communication equipment such that passwords can be exchanged with human troops. If the person gives the wrong password, the robot has the capability to fire at the target using rubber bullets or a swivel-mounted gun.

3 International humanitarian law: three principles

Instead of the pacifist view of world peace, International Humanitarian Law (IHL) acknowledges that wars are part of society and defines Laws of Armed Conflict (LoAC) from a humanitarian perspective. In other words, IHL attempts to balance the human requirements (i.e. right to life) with the necessities of war, and as such works towards the reduction of human suffering in and during wars. Individual moral and legal responsibility plays, since the Nuremberg Trials against Nazi war criminals, an important role in the LoAC and IHL. Prior to the Nuremberg trials, soldiers, in order to be exculpated from wrongdoing, could cite the so-called Superior Orders Defence, which required them only to prove that the orders they carried out were duly authorized. However, the Nuremberg Trials established the so-called Nuremberg Defence. In addition to showing that an order had been duly authorized within the command structure of the military, the Nuremberg Defence sets out two further criteria, the moral perception and moral choice criteria, which, respectively, require soldiers to prove that the actions set out by an order were (1) morally and legally legitimate and (2) unavoidable (May 2005). After Nuremberg, the common excuse of wrongdoers that ‘they were just following orders’ is no longer valid. In summary, the backbone of IHL lies with three principles, being (A) necessity, (B) discrimination and (C) proportionality. First, necessity requires combat forces to engage only in those acts necessary to accomplish a legitimate military objective. Attacks shall be limited strictly to military objectives such as military bases, while hospitals and graveyards are not necessary to attack. The destruction of the Iraqi air defense systems during the Operation Desert Storm serves as a good example of necessity as this led to a strategic advantage in air space (air superiority). Secondly, discrimination, or distinction as it is also referred to, requires discriminating between combatants and non-combatants. Examples of non-combatants are civilians and

civilian property but also include prisoners of war and wounded combatants out of combat. In this respective, tasks like identification and classification are important. Third, proportionality prohibits the use of any kind or degree of force that exceeds that needed to accomplish the military objective. Proportionality is thus a balancing act between the military advantage and the harm inflicted and as such strives to limit collateral damage.

Clearly, those in violation of these three principles may be held liable for war crimes. The LoAC distinguishes justice in the conduct (i.e. *jus in bello*) of war and justice in the declaration of a war (i.e. *jus ad bellum*). Individual soldiers, or combatants, can be held responsible for justice in the conduct of war (*in bello*), whereas state leaders can be held responsible for unjustly declaring a war. In other word, while soldiers cannot be held responsible for participation in an unjust war that violates the criteria of *jus ad bellum*, they can be held responsible for violations of *jus in bello*. Or as Michael Walzer (Walzer 2006) puts it, a soldier is on its own for any crimes committed *during* war.

4 Effect of advance warfare technologies

There is much debate between leading experts of IHL whether today’s deployment of drones is lawful or not (Heyns 2013b; O’Connell 2010). The complexity of the debate on drones is exacerbated due to the advances in autonomous drones and its related campaign against the killer robots (Human Rights Watch 2012). It is important to distinguish between effects due to physical distance and effects resulting from higher levels of autonomy. Regarding the first, physical distance, it is important to understand the effect of physical distance on moral perception: human operators see the battlefield by means of mediating technologies. And with regard to increased autonomy, there is a growing concern that ethical values are violated (Matthias 2004; Sparrow 2007; Human Rights Watch 2012), up to the point that we are losing humanity. It can thus be concluded that both physical distance and increased levels of autonomy changes how information is perceived and therefore affects moral perception. Prior to discussion moral perception, the physical distance and autonomy are discussed in the following two subsections.

4.1 Physical distance to the battlefield

Today, tele-operated, unmanned or remotely piloted vehicles are considered examples of technologies that enlarge the physical distance between the battlefield and the human operators, up to the point that operators are morally disengaged (Royakkers and van Est 2010). However, the discussion regarding technologies that enlarge the distance

to and from the battlefield is much older. Since the introduction of artillery guns, philosophers and ethicists have debated the effect of technology on distance to and from the battlefield. Clearly, the larger physical distance to the battlefield has benefits as well as downsides. First and foremost, the enlarged distance to the battlefield significantly reduces the possibility of death or serious injury amongst service personnel. Secondly, given that they do not face an immediate threat to their safety, the stress soldiers experience in combat is diminished. Stress affects decision-making because it influences how human beings interpret their environment and frame certain issues. To illustrate the point, consider the infamous My Lai massacre that occurred during the Vietnam War. Fearing that the inhabitants of the hamlet of My Lai were Vietcong guerrillas posing as civilians, US soldiers experienced high levels of stress and, as a result framed any information in favour of their fear thereby failing to apply the discrimination criterion accurately. This led to one of the worst massacres in post-war history. A decrease in stress, then, might lead to greater awareness as well as more accurate interpretations of morally relevant facts in a combat situation. Following these arguments, the effects of advanced warfare technologies on decision-making seem more positive, rather than negative. That said, the reduction in stress can also have negative effects. While it is correct that too much stress diminishes human decision-making capacities, stress can have positive effects on an operator's alertness. It has, for example, been demonstrated that under-load conditions negatively impact decision-making and performance (Endsley and Kiris 1995).

Moreover, the category of tele-operated drones has two characteristics that are, compared to the piloted jet fighters or targeting systems, novel. First, drones have *increased surveillance capabilities* due a rich sensor suite. Such a modern sensor suite entails not only more sensors (compared to traditional targeting systems and jet fighters) but also better quality sensors. Secondly, today's drones have a *larger surveillance capacity* due to increased flying time compared to jet fighters or helicopters. Both characteristics lead, at least theoretically, to reduced collateral damage (due to improved situational assessment allowing to time any attack such that the expected collateral damage is least).

There are also a number of downsides reported, especially in the context of drones. First, drones operators have an increased risk of post-traumatic stress as it has been reported that drone pilots bond with objects with the area of interest (Abé 2012). Secondly, soldiers are directly removed from the horrors of war, and there is the risk that operators see the enemy not as humans but as blips on a screen. This introduces the very real danger of losing the deterrent that such horrors provide. This is, by some,

referred to as the game mentality or moral disengagement of drone pilots (Royakkers and van Est 2010; Sharkey 2010). Third, it is worried that operators become trigger happy with remote controlled armaments, situated as they are in complete safety, distant from the conflict zone (cf. incorrect judgement of necessity principle). And fourth, the claim by many states that drones are more precise is debated by many human rights organizations (International Human Rights and Conflict Resolution Clinic 2012).

4.2 Increased autonomy/levels of autonomy or automation

Many military organizations consider equipping drones with high levels of autonomy in order to reduce the degree and amount of interactions between the machine and its operators, thereby lowering the operational cost. However, there is growing concern that, as a result from these high levels of autonomy, ethical values are under pressure (Cummings 2003; Royakkers and van Est 2010; Sharkey 2010) up to the point that one or more moral values cross a critical threshold. For example, placing the human operator outside of the decision-making loop, Human Right Watch (2012) argues, opens up a responsibility gap—that is, an ambiguity with respect to whom should be hold responsible for actions in case of misdoing (Matthias 2004; Sparrow 2007). Is it the robot, is it the programmer or is it the army?

But we believe that the discussion on the responsibility gap is subtler. Consequently, we discuss two items. First, even when a drone operates a mission in full autonomous mode—that is that the drone achieves a military objective without the human input or control—someone within the chain of command has made the decision to deploy the autonomous drone for the particular operation and thereby has responsibility. Second, between full autonomous drones and tele-operated drones lies an opportunity space of interaction modes, which are (in the discipline of human factors) known as levels of automation. Moreover, the notion of autonomy seems to be used differently across disciplines, leading to ambiguity what it means to be autonomous. Philosophers define moral autonomy as a capability to (A) pursue the concept of good life (cf. express free will) and (B) develop desires about our desires (cf. second-order desires). The latter allows humans to distance ourselves from our initial believes and critically reflect upon them. By contrast, the discipline of cognitive engineering and human factors interprets autonomy in the context of human machine interaction, which we will refer to as operational autonomy. Bradshaw et al. (2004) claim that autonomy has two senses, namely self-sufficiency and self-directedness. The first, self-sufficiency, relates to the set of actions that are possible with or without the help of another actor, whereas the second, self-directedness, relates

to the set of actions that are permitted and obliged. Low self-directedness indicates that, although potentially capable of performing the task, the agent is not permitted to do so, whereas high self-directedness indicates that the agent has authority over its own actions, though it does not necessarily imply sufficient competence. Ron Arkin prefers to talk about constraints instead of permitted and obliged actions as the laws of armed conflict constrains the deployment of force rather than obliged the use of force (Arkin 2009). For instance, Rules of Engagement provide explicit criteria for the use of force as well as for the utilization of certain components of weapons systems. A pilot of a fighter jet might be constrained to use its fire radar sensor to lock on enemy or suspect air tracks. The reason for this prohibition is that opposing forces could interpret this as a hostile intention unnecessary risking use of force and life. Likewise, a drone might be restricted in its use of specific cameras due to privacy regulations.

The self-sufficiency axis of autonomy matches nicely with the level of automation taxonomies proposed in the field of human factors. Various studies have been conducted that provide indications on the level of control that can be allocated towards a human or a machine. These studies aim to find the sweet spot between full autonomy and manual operations given the use of the technology in the context of deployment. Such intermediate levels, often denoted as mixed initiative, shared control or adaptive/adaptable automation, are seen as the most promising area for improved efficiency in interactive systems. Prior to such intermediate levels, Fitts stated that that humans and computers have different capabilities by composing a list of general task abilities summarizing where “Men-Are-Better-At” and where “Machines-Are-Better-At”. The so-called Fitts’ list (1951) helped designers at that time to allocate functions or tasks to either a human or a machine. However, since Fitts a number of taxonomies are proposed which can, roughly, be summarized as an evaluation of three generations. The first generation is based upon the seminal work of Sheridan and Verplank (1978), who were among the first to acknowledge that automation is not an ‘*all or nothing*’ fashion (thus contrasting Fitts’ assumption). Sheridan and Verplank defined 10 levels of automation that differentiate in the actions by the machine and the information offered (by the machine) to the human. The second generation is based upon the work of Parasuraman et al. (2000). In this model, information processing is divided into four stages alike how a human processes information. The model claims that all of these stages should be automated at a different level of automation based on primary (e.g. human performance) and secondary (e.g. automation reliability) criteria. The third generation builds upon the four-stage model of Parasuraman, Sheridan and Wickens but addresses the question how to divide the

work within each of the four stages (Arciszewski et al. 2009). The model recognizes only five different levels of automation that are clearly differentiated by (A) whether the machine or the human engages in action and (B) whether the machine actively updates the human. Furthermore, the model of Arciszewski and de Greef focuses on the central domain objects and relevant attributes and leads to a fine-grained division of labour between the human and machine with regard to the task. Working agreements, which allow making explicit which tasks and objects are in control of the human or the machine, facilitates such a fine-grained division of labour. Consider, as an example, a human who is responsible for identifying track objects (e.g. airplanes, vessels) in an area of interest. While some objects are unambiguous to identify (i.e. a radar installation) given the characteristics of the object, others lack such clear characteristics, increasing the ambiguity in terms of identification. Using working agreements, the operator, for instance, may delegate the task of identifying the unambiguous objects to the machine, while remaining in control of the more cognitive demanding ambiguous objects.

Depending on the interface design, increased levels of automaton affects how information is offered to the human. On the one hand, an operationally autonomous machine may process and filter large amounts of information before passing on selected information to an operator. On the other hand, such operational autonomous machines might supply operators with too much information. There is a whole field of interface design challenged by the balance between filtering information and supplying all information (cf. Vicente and Rasmussen 1992). Keep in mind that the problem is worsened due to the fact that airborne drones can have many sensors and can remain in the air for long periods of time. Processing the amount of information they provide may be difficult for a single operator. In both cases—the undersupply and oversupply of information—it becomes difficult for operators to filter out morally relevant facts, hereafter discussed as moral perception.

5 Moral perception affected and some requirements

The concept of moral perception refers to the knowledge of the morally relevant facts in a particular situation. Since Nuremberg, in order to be exculpated from wrongdoing, an individual has to prove that he could not have acquired knowledge of the morally relevant facts. They must prove that, based on the information they had at the time, they thought an order was legitimate. Knowledge of morally relevant facts enables soldiers to apply the key principles of discrimination, necessity and proportionality in order to assess an order.

However, in light of the rise of advanced warfare technologies, the way in which soldiers acquire knowledge of relevant moral facts is being transformed. Depending on the context, it might be more difficult to hold operators culpable for, say, applying force to a target. One potential reason for this is that in case force is applied to the wrong target, operators could argue that due to the restrictions imposed by the technology, they did not have full situational awareness and should therefore be exculpated from any wrongdoing. As a result, the deployment of NCTs may undermine a commitment to individual responsibility. However, following the design practise of value sensitive design (De Greef et al. 2013; Friedman et al. 2003), we believe that *design for responsibility* is the way forward, also in the use-case of increased levels of autonomy.

This leads to two immediate requirements. Firstly and from a more technologically oriented perspective, engineers designing military equipment must be sensitive to how different types of technology impact on the moral perception of their operators. That is to say, they must take into account how psychological factors impact information processing and shape the perception of morally relevant facts. Secondly and from a more legally and normatively oriented perspective, advanced warfare technologies must be designed in order to minimize any distortions or unnecessary restrictions of their operators' moral perception. Overall, sound design must always be design that enhances, rather than undermines, the preconditions for individual responsibility. Ensuring this is, in our view, one of the central moral obligations of engineers and designers. In the next part of this paper, we discuss the partnership approach and provide a soft architecture that explains the commitment to responsibility.

6 Partnerships and moral preception

How can we ensure that individuals perceive the relevant facts in a give situation? One position holds that we can't. In fact, it contends that if individuals are unreliable decision-makers due to stress factors and human information processing limitations, it might be a good idea to take them out of any decision-making loop altogether. In this case, machines are made fully operationally autonomous. This solution is proposed by the US roboticist Ronald Arkin (2009). Although Arkin expresses reasonable worries, taking the human completely out of the loop leads to the underutilization of human experiences and capabilities. It remains a fact that specific humans capabilities are, at the time of writing, hard to replicate by artificial intelligence. While, for example, computers are really strong in executing many calculations on large datasets consistently, they lack creativity or a capability to recognize patterns

that humans recognize quite easily. To do so, humans use a variety of psychological tools such as the knowledge-based reasoning mechanism (Rasmussen 1986). The knowledge-based reasoning mechanism allows humans to cope with novel and unexpected situations by using basic fundamental knowledge (e.g. principles, physical laws) that governs the specific domain. Today's artificial intelligence technologies fail to model precisely the set of tools that allow being creative and recognize patterns. It is, for instance, difficult to see how a contemporary machine could interpret complex behaviour characterizing hostile activities. However, a computer is at times capable to recognize less complex but consistent patterns. Up to know, many claim that a computer is incapable to distinguishing a combatant from non-combatant merely because human behaviour is complex. However, classifying particular objects (by combining databases or recognizing particular shapes) would typically be something that machines could outperform humans. So, while Arkin is right to point out that humans are bad interpreting complex information under stressful conditions, machines presently lack the reasoning capacities that allow interpreting complex behavioural patterns to the same extend humans without stress would interpret this.

Faced with this problem, it is a commonsensical response, we think, to try and 'team up' humans and artificial agents, into a partnership (cf. Klein et al. 2004). The partnership approach follows the joint cognitive systems (Hollnagel and Woods 2005) paradigm shift from *automation extending* human capabilities to *automation partnering with* the human. The artificial agent partners with the human and as such collaborates in a symbiotic fashion to achieve the best performance while operating within safety boundaries. Stated differently, it stresses that neither the machine nor the human is able to solve problems individually, but that both are partners in solving problems effectively and efficiently. Alike human partnerships, explicit agreements are made and mutual reciprocity exists between the partners meaning that you need each other to achieve goals and engage together in tasks and activities. Some scholars regard artificial agents in partnership as *members of a human-machine team* (cf. Salas et al. 2008, p. 544) or *e-partners*. The latter are proposed in various domains such as space missions (Neerinx and Grant 2010), self-health care services (Blanson Henkemans 2009) or navy operations (de Greef 2012).

Within the joint cognitive systems paradigm, the collaboration between the human and the machine centralizes around the joint activity of these actors and their joint performance (see Fig. 1, left). The concept of joint activity is defined as an activity "*that is carried out by an ensemble of people acting in coordination with each other*" (Clark

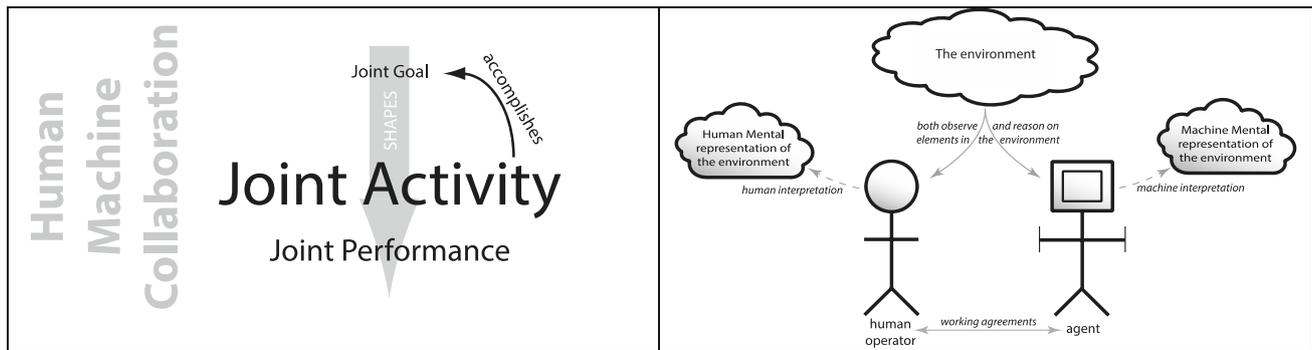


Fig. 1 *Left:* a human and machine collaborate on joint activities leading to a joint performance. The joint activities accomplish a joint goal. *Right:* the proposed architecture for the partnership paradigm

1996, p. 4). While Clark studied human teams, a shift from human teams to human–machine teams is easily facilitated. Like humans are goal driven and require activities to achieve those goals, ensembles have a joint goal and require joint activities in order to accomplish these goals. Joint activity leads to joint goal accomplishment and is measurable as joint performance.

The right side of Fig. 1 presents the architecture for the partnership paradigm and reveals two distinctive features. First, instead of letting an activity be performed by either the human, the machine or a combination of both, a mechanism is put forward for whereby both parties do their job *concurrently*. In this way, each actor arrives at an own interpretation of the world thereby constructing a human representation of the world and a machine representation of the world *at the same time*. Each can deposit the information pertaining to their view of the world in their respective ‘storage space’, where the results of their respective computational and cognitive efforts can be written to. Having these two distinct views on the world is akin to two people coming to different conclusions based on the same set of (non-conclusive) data. The machine may be viewed as a team member with a limited but scrupulously objective view of the world.

It is thus assumed that the machine is to some extent capable of deriving its own interpretation of the situation and has reasoning capability. Both views can be compared for differences and argumentations. One important aspect of this concept is the fact that the machine always calculates (cf. reasons) its view, independent of whether the user is dealing with the same or not.

Secondly, working agreements are explicit contracts between the human and the machine about the division of work (Arciszewski et al. 2009; Miller and Parasuraman 2007; Parasuraman and Miller 2004). Having working agreements minimizes the automation-human coordination asymmetry (Woods et al. 2004) because working agreements define an a priori explicit contract what and what not

to delegated to the automation. Working agreements can reflect easily the opinions of the system designers and users on the proficiency of the software and can also serve establishing trust in the workings of the machine by starting at fairly low levels of automation. They can also mirror thoughts on how much work is ultimately delegated to the machine under the political and strategic constraints of the mission. The considerations that are taken into account when drafting these working agreements are the same as the secondary evaluation criteria brought to the fore by Parasuraman et al. (2000): reliability and the costs of actions.

The concept of partnerships and working agreements was tested and evaluated in a naval military context (see de Greef et al. 2010). The eight navy officers who participated in the study highly appreciated the division of labour between human and machines introduced by the working agreement, especially when decisions had to be made under pressure. Some navy officers claimed that they didn’t fully agree with the machine but that they accepted, understood and respected the interpretation of the machine. The officers liked the working agreements and were relieved that they could focus on the more demanding tasks while having the machine carry out relatively easy tasks.

In the light of these findings, the effect of the proposed partnership architecture on moral perception is potentially positive. Firstly, the partnership approach increases efficiency via a user-inspired division of labour, lowering the stress experienced by those operating them. Importantly, the operator remains involved in the decision-making process, especially for those tasks and objects that are deemed important and cognitive demanding. Secondly, the partnership proposal introduces an interesting dynamic between human operators and their machines. Just as, during ordinary teamwork, human team partners may develop different perspectives on a situation, machines and humans may develop different perspectives on a situation. This can be taken to our advantage. Operators can use the

perspective provided by their machine to check whether they are missing morally relevant facts. The machine may even flag up aspects of a situation that the operator might have otherwise overlooked. This safety mechanism challenges operators to think critically about actions and how these fit within the moral and legal framework.

If these points are sound, the proposed partnership architecture can protect a commitment to responsibility within the armed forces. First, operators will be responsible for the terms of their working agreements with their machine. This raises issues about foresight, negligence and so on that we cannot tackle here. For now, it suffices to note that the operator remains firmly control of his machine—even if there is a physical distance between them or that the machines operates at increased levels of automation. Secondly, working agreements ensure that operators receive the morally relevant facts needed to make decisions that comply with IHL, as well as key moral principles.

7 Conclusion

We started by noting that some commentators have argued in favour of the deployment of advanced warfare technologies during combat. Indeed, there are some benefits associated with such advanced technologies. But the downsides posed by these systems must not be neglected either. Before advanced warfare technologies can be deployed, we need to rest assured that their usage is safe and that they enhance, rather than undermine, human decision-making capacities and a commitment to responsibility. This is important in any type of armed conflict. During military missions, the operational requirements upon advanced warfare technologies and those who operate them are high. Operators and their superiors need reliable information about the complex environment they operate in, especially when they are not directly present.

Following value sensitive design rationales, it is important to (re)design advance warfare technologies for a commitment to responsibility and for that, it is important to discuss ethical values like proportionality, necessity and discrimination. We therefore propose the partnership architecture that embeds concurrent views of the world and we discuss working agreements. Both ensure that operators use appropriate information in the decision-making process. There are important choices to be made, and sound design is ‘design for responsibility’. The partnership approach is a promising way forward, especially in the light of a commitment to responsibility and when compared to proposals for fully operationally autonomous machines.

Acknowledgments This work is part of the research Project 313-99-260 ‘Military Human Enhancement: design for responsibility

in combat systems’ that is financed by the Netherlands Organization for Scientific Research (NWO) in the program line ‘Societal Responsible Innovation’. This journal paper is an extended version of the paper ‘New Combat Technologies: safeguarding moral perception and responsibility via e-Partners’ that appeared in the proceedings of the 11th International Conference on Naturalistic Decision Making (NDM 2013) edited by H. Chaudet, L. Pellegrin & N. Bonnardel (ISBN 979-10-92329-00-1). The conference took place in Marseille in France on 21–24 May 2013 and was published by Arpege Science Publishing, Paris, France.

Open Access This article is distributed under the terms of the Creative Commons Attribution License which permits any use, distribution, and reproduction in any medium, provided the original author(s) and the source are credited.

References

- Abé N (2012) Dreams in infrared: the woes of an American drone operator (Part 2: An Unpopular Job). in *Der Spiegel*, issue 50
- Altman A, Wellman CH (2009) A liberal theory of international justice. *Annals of Physics* (vol 2, pp 1–42). Oxford University Press. doi:10.1093/acprof
- Arciszewski HFR, de Greef TE, van Delft JH (2009) Adaptive automation in a naval combat management system. *IEEE Trans Syst Man Cybern A Syst Hum* 39(6):1188–1199
- Arkin R (2009) *Governing lethal behaviour in autonomous robots*. Taylor & Francis, Baco Raton
- BEASystems (2011) *Taranis—looking to the future*. BEA Syst. Retrieved from 23 Oct 2013. http://www.baesystems.com/cs/groups/public/documents/document/mdaw/mdy2/~edis/baes_057655.pdf
- Blanson Henkemans OA (2009) ePartner for self-care/how to enhance ehealth with personal computer assistants. *Fac EWI Dep ManMachine Interaction Group*. Delft University of Technology, Delft
- Bradshaw JM, Feltovich P, Jung H, Kulkarni S, Taysom W, Uszok A (2004) Dimensions of adjustable autonomy and mixed-initiative interaction. In: Carbonell JG, Siekmann J (eds) *Agents and computational autonomy: potential, risks, and solutions*. Lecture notes in artificial intelligence, vol 2969. Springer, pp 17–39
- Clark H (1996) *Using language*. Cambridge University Press, Cambridge
- Cummings ML (2003) Automation and accountability in decision support system interface design. *J Technol Stud* 32:23–31
- De Greef T (2012) *ePartners for dynamic task allocation and coordination*. Ph.D. dissertation: Delft
- De Greef TE, Arciszewski HFR, Neerinx MA (2010) Adaptive automation based on an object-oriented task model: implementation and evaluation in a realistic C2 environment. *J Cogn Eng Decis Mak* 31:152–182
- De Greef T, Mohabir A, van der Poel I, Neerinx MA (2013) sCEthics: embedding ethical values in cognitive engineering. In proceedings of the 31st European conference on cognitive ergonomics
- Endsley M, Kiris E (1995) The out-of-the-loop performance problem and level of control in automation. *Hum Factors* 37:381–394
- Friedman B, Kahn PH, Borning A (2003) Value sensitive design and information systems. In: Zhang P, Galletta D (eds) *Technology* 3(6):1–27
- Heyns C (2013a) Report of the special rapporteur on extrajudicial, summary or arbitrary executions (A/HRC/23/47)
- Heyns C (2013b) report of the special rapporteur on extrajudicial, summary or arbitrary executions (A/68/382)

- Hollnagel E, Woods DD (2005) *Joint cognitive systems: foundations of cognitive systems engineering*. Taylor & Francis, Boca Raton
- Human Rights Watch (2012) *Losing humanity: the case against killer robots*. Human Rights Watch, New York
- Ignatieff M (2000) *Virtual wars: Kosovo and beyond*. Vintage, London
- International Human Rights and Conflict Resolution Clinic (2012) *Living under drones: death, injury, and trauma to civilians from US drone practices in Pakistan*. Stanford Law School and NYU School of Law
- Klein G, Woods DD, Bradshaw JM, Hoffman RR, Feltovich PJ (2004) Ten challenges for marking automation a “Team Player” in joint human-agent activity. *IEEE Intell Syst* 19(6):91–95
- Matthias A (2004) The responsibility gap. Ascribing responsibility for the actions of learning automata. *Ethics Inf Technol* 6:175–183
- May, L. (2005). Crimes against humanity: a normative account. *Eur J Int Law* vol 18, p 310. Cambridge University Press
- Miller CA, Parasuraman R (2007) Designing for flexible interaction between humans and automation: delegation interfaces for supervisory control. *Hum Factors* 49(1):57–75
- Neerinx MA, Grant T (2010) Evolution of electronics partners: human-automation operations and ePartners during planetary missions. *J Cosmol* 12:3825–3833
- O’Connell ME (2010) *Unlawful killing with combat drones: a case study of Pakistan, 2004–2009*. New America A Review, pp 2004–2009
- Parasuraman R, Miller CA (2004) Trust and etiquette in high-criticality automated systems. *Commun ACM* 47(4):51–55
- Parasuraman R, Sheridan TB, Wickens CD (2000) A model for types and levels of human interaction with automation. *IEEE Trans Syst Man Cybern A Syst Hum* 30(3):286–297
- Rasmussen J (1986) *Information processing and human-machine interaction: an approach to cognitive engineering*. North-Holland, Amsterdam
- Royakkers L, van Est R (2010) The cubicle warrior: the marionette of digitalized warfare. *Ethics Inf Technol* 12(3):289–296
- Salas E, Cooke N, Rosen M (2008) On teams, teamwork, and team performance: discoveries and developments. *Hum Factors* 50(3):540–547
- Sharkey N (2010) Saying “No!” to lethal autonomous targeting. *J Mil Ethics* 9(4):369–383. doi:10.1080/15027570.2010.537903
- Shaw M (2005) *The new western way of war: risk-transfer war and its crisis in Iraq*. Sierra pp VII, 183. Polity Press. Retrieved from. <http://eprints.sussex.ac.uk/1506/>
- Sheridan TB, Verplank WL (1978) *Human and computer control of undersea teleoperators*. DTIC Res
- Sparrow R (2007) Killer robots. *J Appl Philos* 24(1):62–77
- Strawser BJ (2010) Moral predators: the duty to employ uninhabited aerial vehicles. *J Mil Ethics* 9(4):342–368
- Strawser BJ (2013) *Killing by remote: ethics of an unmanned military*. Oxford University Press, Oxford
- Vicente KJ, Rasmussen J (1992) Ecological interface design: theoretical foundations. *IEEE Trans Syst Man Cybern* 22(4):589–606
- Walzer M (2006) *Just and unjust wars: a moral argument with historical illustrations*, vol 32. Basic Books, New York, p 361
- Woods DD, Tittle J, Feil M, Roesler A (2004) Envisioning human-robot coordination in future operations. *IEEE Trans Syst Man Cybern C Appl Rev* 34(2):210–218