Towards a principle for the human supervisory control of robot weapons*

Noel Sharkey
University of Sheffield, UK

Abstract:
The aim is to explore the delicate balance between human reasoning and the legal requirements for the appropriate supervisory control of robot weapons. We start by examining the limitations of automatic versus aided target recognition and then review some of the relevant psychological literature on reasoning. A new framework is introduced that reframes autonomy/semi-autonomy in terms of levels of supervisory control. This allows for greater transparency in command and control and the allocation of responsibility. Finally, so-called, human supervised autonomy is assessed in terms of the supervisory control framework.

Keywords: autonomous weapons, supervisory control, robot weapons, international humanitarian law, reasoning

In a world where computing is taking us to new levels of automation, we must ensure that the decision to kill remains firmly under human control. Most new technological artefacts are controlled by computer chips and the technologies of violence are no exception: computer devices are becoming ubiquitous for most modern weapons and guidance control systems. Currently almost all of these weapons are under "supervisory control", where a computer program mediates human control.[1]

Humans need to exercise meaningful control over weapons systems[2] to counter many of the problems that arise from automation. The United States Department of Defense points out a number of the potential difficulties with entirely computerising robot weapon: human error, human-machine interaction failures, malfunctions, communications degradation, software coding errors, enemy cyber attacks, infiltration into the industrial supply chain,

* Thank you to Maya Brehm, Article 36 and Dr. Amanda Sharkey, Computer Science, University of Sheffield for helpful comments on earlier drafts of this article.

[1] Supervisory control stands in contrast to direct human control such as aiming a conventional rifle and pressing the trigger manually.

[2] «The exercise of control over the use of weapons, and, concomitant responsibility and accountability for consequences are fundamental to the governance of the use of force and to the protection of the human person». Article 36 (2013) Memorandum for delegates to the Convention on Certain Conventional Weapons (CCW). Downloaded from: http://bitly.com/NJoVG3. Last accessed March 2 2014.

jamming, spoofing, decoys, other enemy countermeasures or actions, unanticipated situations on the battlefield.[3]

Even though such difficulties are well known, there is ever-increasing push by states to develop autonomous robot weapons that could move outside the reach of human supervisory control. The US has conducted advanced testing on a number of autonomous weapons platforms such as X-47b – a fast subsonic autonomous jet that can now take off and land on aircraft carriers, the Crusher – a 7 ton autonomous ground robot, and an autonomous hunting submarine. The Chinese are working on the Anjain supersonic autonomous air-to-air combat vehicle. The Russians are developing an autonomous Skat jet fighter. Israel has the autonomous Guardium ground robot and the UK is in advanced testing of the Taranis – a fully autonomous intercontinental combat aircraft.

Currently, public reports on the testing of these robot devices have only been about the weapons-carrying platforms and not the weapons systems as a whole. Although autonomous robot platforms can have multiple useful purposes, arming them to select targets and attack them without deliberative human supervision raises serious legal, ethical, technical and international security concerns. The main concern of this paper is to explain the psychological underpinnings and problems with human supervisory control and to emphasise the need for a comprehensive method to ensure that appropriate deliberative human reasoning is utilised.

1. Automatic versus aided target recognition

A major problem with weapons systems in which a computer program selects targets and initiates attack is that to identify and select targets requires well-defined target recognition software. Yet current automatic target recognition methods used by the military are not fit for purpose except in narrowly restricted and highly uncluttered environments. Some of examples of the main methods are:

1. *Shape detection* makes it possible to recognise a tank in an uncluttered environment, such as a sandy desert plain. Medium to high cluttered environments introduce an unacceptably high false alarm rate. It has proved extremely difficult to distinguish between a truck and a tank or any vehicle amongst clutter, such as

[3] US Department of Defense, *Autonomy in Weapon Systems*, Directive 3000.09, November 21 2012.

trees.[4] For example, such systems use feature detection that would have difficulties distinguishing between a smooth overhanging branch and the barrel of a large gun.

2. *Thermal imaging* detects heat radiating from an object and shows its movement. But it would be difficult for an autonomous system to parse the image and guarantee that it can distinguish between a tank and a school bus. And certainly could not be used to distinguish between a combatant and civilians.

3. *Radiation detection* is used by loitering munitions, such as the Israeli Harpy, to detect radar signals and determine if they are friendly. If not then the Harpy dive bombs the radar. It is assumed that the radar is part of an anti-aircraft installation, otherwise radar detection doesn't have any other means to determine the legitimacy of a target.

4. *Acoustic Direction Finding* is a method of using minute differences between the times that sound reaches two or more separated microphones to calculate the location of the sound. For example, this method is used by the Red Owl sniper detection system that sits on top of an iRobot Packbot. It uses 4 microphones to determine the location of a gunshot and then shines a laser designator and points other sensing equipment at it. A problem here is that other acoustic effects such as fireworks, ricochet and friendly fire may be detected and responded to. A human is need to exercise judgement about the legitimacy of the target.

Despite decades of research, the limitations of these targeting methods are severe. The idea of developing autonomous weapons, outside of narrow restrictions, that could comply with the legal requirements on the use of force under international human rights law and international humanitarian law is entirely speculative and cannot be guaranteed. This is why nations like the US and the UK have made it clear that for the time being there will always be a human in the loop for lethality decisions.[5, 6, 7]

[4] J.A. Raches, «Review of current aided/automatic target acquisition technology for military target acquisition tasks», *Optical Engineering*, n. 50 (2011), 7, pp. 1-8.

[5] The technical problems may or may not be alleviated by unanticipated scientific breakthroughs at some unspecified point in the future. However, it is a very risky strategy to rely on conjecture. Billions of dollars of investment in research cannot guarantee success. All we can say is that at present, and into the foreseeable future, the use of autonomous weapons in most circumstances would violate the Principles of Distinction and Proportionality (see: N. Sharkey, «The Evitability of Autonomous Robot Warfare», *International Review of the Red Cross*, n. 94 (2012), pp. 787-799; and N. Sharkey, «Saying – No! to Lethal Autonomous Targeting», *Journal of Military Ethics*, n. 4 (2010), 9, pp. 299-313.

[6] Concerns have also been expressed about how combating algorithms could serious disrupt international security and trigger unintended wars. N. Sharkey, «The Automation and Proliferation of Military Drones and the Protection of Civilians», *Journal of Law, Innovation and Technology,* n. 3 (2011), 2 pp. 229-240.

When the US Department of Defence issued the first policy document on autonomous weapons, they stated: «Autonomous and semi-autonomous weapon systems shall be designed to allow commanders and operators to exercise appropriate levels of human judgment over the use of force».[8]

On 26 March 2013, the Parliamentary Under Secretary of State, Lord Astor of Hever, replied to questioning about autonomous weapons at a House of Lords debate. He said: «the MoD currently has no intention of developing systems that operate without human intervention». And also that: «Fully autonomous systems rely on a certain level of artificial intelligence for making high-level decisions from a very complex environmental input, the result of which might not be fully predictable at a very detailed level. However, let us be absolutely clear that the operation of weapons systems will always be under human control».[9]

Even an improved ability to recognise targets does not enable machines to assess whether a target is legitimate and whether the attack as a whole is permissible. The appropriateness and legality of an attack is context-dependent and needs to be assessed on a case-by-case basis. Thus a more appropriate use of these methods would be to assist human supervisory control to achieve more precise and accurate targeting by a human exerting the power of deliberative reasoning and judgement.

With increasing computerised weapons systems, we need to ensure that the partnership between human and machine is more humanitarian than machines or humans operating alone. We can also exploit the human-machine relationship to ensure greater accountability by using data traces of actions.

2. The delicate human and computer balancing act

The control of weapons mediated by computer programs raises its own problems. Perhaps the most important of these is the delicate human-computer balancing act. Because humans sometimes fail at some tasks, it does not mean that machines can do them any better. It can simply mean that humans are being asked to perform in a mode of operation that is not well suited to human psychology. This needs to be part of the equation of

[7] Perhaps an even stronger case is that regardless of how much the technology progresses, the very idea of delegating the decision to kill to a machine crosses a fundamental moral line. See: P. Asaro, «On Banning Autonomous Weapon Systems: Human Rights, Automation and the Dehumanisation of Lethal Decision-Making», *International Review of the Red Cross*, n. 94 (2012), pp. 687-709; and C. Heyns, «Report of the special Rapporteur on extrajudicial summary or arbitrary executions», to the *Human Rights Council Twenty-third session* (2013). Downloaded from: http://bitly.com/1hM9GXr. Last accessed March 5 2014.

[8] US Department of Defense, *Autonomy in Weapon Systems*, cit.

[9] Downloaded from: http://www.publications.parliament.uk/pa/ld201213/ldhansrd/text/130326-0001.htm#st_14. Last accessed March 3 2014.

ensuring efficient and meaningful human supervisory control of weapons.

Computers are better and more efficient at some task than humans while humans are better at other tasks. Examples are provided in Table 1.

Table 1: examples of tasks computers and humans are better at[10]

Computers	Humans
calculating numbers	deliberative reasoning
searching large data sets	perceiving patterns
responding quickly to control tasks	meta-cognition
performing repetitive routine tasks simultaneously	reasoning inductively
carrying out multiple complex tasks	applying diverse experience to novel tasks
sorting data	exercising meaningful judgment

If we get the balancing act right we could have more precision and accurate targeting with less collateral damage and better predictable compliance with International Humanitarian Law. But getting it wrong could result in considerable humanitarian problems.

To say that a human is in-the-loop does not clarify the degree of human involvement. Being in the loop could be a simple as pressing a button that lights up to say that a target has been detected or it could mean exercising full human judgement about the legitimacy of a target before initiating an attack. Figure 1 lists five levels for the human supervisory control of weapons.

In order to understand the strengths and limitations of each of the methods, it is important to understand the underlying psychology of the decision process that a human operator can employ in each of the levels of control. Designing a supervisory control system to enable a human to operate weapons must take into account the psychological limits of decision making under different conditions.

Daniel Kahneman, who won the Nobel Memorial prize for his work on reasoning, describes human decision making in terms of two types of psychological processes that he terms System 1 and System 2.[11] The first refers to fast processes that are always cued automatically and the second refers to slower controlled and deliberative processes. For example, as children we are trained to automatically (quickly) answer simple arithmetic questions such as 2x2. But when we see a problem such as 36x241 we need to slow down and deliberatively calculate the answer.

There is, of course, debate in the psychological literature about experimental details,

[10] See also M.L. Cummings, «Automation Bias in Intelligent Time Critical Decisions Support Systems», *American Institute of Aeronautics and Astronautics*, AIAA 3rd Intelligent Systems Conference Chicago, 2004.
[11] D. Kahneman, *Thinking, Fast and Slow*, Penguin Books, London 2011.

the underlying brain mechanisms[12] and whether it is possible to create a unified model of the dual processes.[13] However, the distinction between automatic and controlled processes is well established and follows from more than 100 years of research on dual processing starting with William James (1890).[14] There is considerable and substantial agreement that automatic and deliberative processes characterise the nature of human reasoning. The main aim here is to extract general properties for the two types with regards to the different decision behaviours that can result when controlling weapons.

Kahneman characterises deliberative reasoning as your conscious self-it is the thinking you.[15] The deliberative processes always come into play after the automatic but, in Kahneman's words, they are lazy. They will go along with the automatic processes unless there is something surprising or irregular and/or we are operating in novel circumstances or performing tasks that require vigilance and/or deliberation. Deliberative processes require attention and free memory space. Automatic processes dominate if we are performing a task that requires attention or memory and a new task needs to be performed in competition. In fact anything that impacts on memory capacity or attention such as stress or distractions could incapacitate deliberative reasoning.

Automatic processes do not require active control or attention. Normally both systems operate seamless together. To illustrate the distinction between the two types we use two of Kahneman's examples.[16]

Example 1

You should quickly state the first answer that comes to mind.

A bat and ball costs $1.10; the bat costs one dollar more than the ball; how much did the ball cost?

The first solution that comes to mind with the automatic process is probably 10c. But with deliberation, you will realise that if the ball costs 10c, the bat will cost a dollar and thus the bat costs only 90c more than the ball. The correct answer is that the ball cost 5c.

[12] For an overview see W. Schneider and J.M. Chen, «Controlled and Automatic Processing: Behavior, Theory and Biological Mechanisms», *Cognitive Science*, n. 27 (2003), pp. 525-559.
[13] For an overview of the issues see S.B.T. Evans and K.E. Stanovich, «Dual-Process Theories of Higher Cognition: Advancing the Debate», *Perspectives on Psychological Science*, n. 8 (2013), 3, pp. 223-241.
[14] W. James, *The Principles of Psychology*, vol. 1, Holt, New York 1890.
[15] D. Kahneman, *Thinking, Fast and Slow*, cit.
[16] *Ibid*.

Example 2

Task. 1: go down each column in turn and whisper to yourself whether each word is printed in lower or upper case. Task 2: Repeat the exercise but this time whisper whether each word is to the right or left of its column.	
LEFT left right RIGHT RIGHT left LEFT right	upper lower LOWER upper UPPER lower LOWER upper

This example illustrates the separation of automatic and deliberative processes by showing them in conflict with one another. The task requires deliberative reasoning because whispering upper/lower or right/left is unusual when reading columns of words. But you will find that one column was significantly easier than the other and the easy column was different for both tasks. This is because we cannot help but automatically read the actual words and this interferes with the deliberative processes.

The relevance to weapons control is that both reasoning types have different properties. The advantage of automatic decision processes is that they can be trained through repetition and practice on routine tasks. They are needed for fast reaction in sports, and for riding a bicycle, driving a car or in military routines. In fact, automaticity is used anytime for routine decisions that have to be made rapidly for predictable events. It works well in an environment that contains useful cues that, via practice, have been (over) rehearsed.

For the right tasks, automatic reasoning can be more optimal than deliberative reasoning and it is not inherently bad. It is necessary for everyday life and works hand in hand with deliberative reasoning. When initiated by well-practiced cues, it reduces much of the tedium in our lives and saves us from a life of indecision. Sharkey and Mitchell (1985)[17] demonstrated that when we read about routine events such as eating in a restaurant or going to a child's birthday party, an appropriate script[18] or culturally specific schema[19] is

[17] N. Sharkey and D.C. Mitchell, «Word Recognition in a Functional Context: the Use of Scripts in Reading», *Journal of Memory and Language*, n. 24 (1985), pp. 253-270.

[18] R. Schank and R. Abelson, *Scripts, Plans, Goals, and Understanding: An Inquiry into Human*

automatically loaded in the background to provide background knowledge for default reasoning. This saves us from having to work out every inference from first principles. For example, when we read «Tommy went to a children's birthday party and little Jenny blew out the *blank*», we can automatically infer that the *blank* stands for candles and we do not have to think about why Jenny blew out the candles or the fact that it must have been her birthday party.

We must also have such scripts for warfare and action on the battlefield. For example read the sentence: «When Harry was on patrol in Afghanistan, he heard a loud *blank* and woke up in hospital with an amputated leg». We automatically infer that the *blank* stands for *explosion* and that it was not a bang from a backfiring car. We also automatically infer the causal link between the explosion and the missing leg. If the *blank* was replaced by *trumpet*, deliberative reasoning would be required to infer a link between a trumpet and an amputated leg.

Members of the armed forces will have rehearsed and over-trained in many routine tasks that require automatic action on order. Fast automatic response can be trained with well practiced cues. These can be useful in military contexts such as when someone shouts "fire in the hole" – a warning that should prompt those hearing it to immediately take cover. The question to ask about automatic reasoning is, does a given domain afford enough regularity to be learnable as an automatic process? When it comes to human supervised targeting, the unpredictable and unanticipated circumstances in a dynamically changing environment play to the weakness of automatic reasoning.

Four of the properties of automatic reasoning from Kahneman *ibid* illustrate how it would be problematic for the supervisory control of weapons. Automatic reasoning:

1. **Neglects ambiguity and suppresses doubt**: automatic processes are all about jumping to conclusions. They are guided by experience. An unambiguous answer pops up immediately and does not allow doubt. Automatic reasoning does not search for alternative interpretations and does not examine uncertainty. So if something looks like it might be a legitimate target in ambiguous circumstances, automatic reasoning will be certain that it is legitimate.

2. **Infers and invents causes and intentions:** automatic reasoning is adept at finding a coherent causal story to link together fragments of available information. Events

Knowledge Structure, Lawrence Erlbaum Associates, Hillsdale NJ 1977.
[19] F.C. Bartlett, *Remembering: A Study in Experimental and Social Psychology*, Cambridge University Press, Cambridge 1932.

including people (or even inanimate objects such as robots) are automatically attributed with intentions that fit the causal story. For example, if a human operator is seeking out patterns of behaviour to determine a lethal drone strike, then seeing people load bales of hay or shovels onto a truck could initiate a causal story that they were loading rifles for an attack. This relates to *assimilation bias* in the human supervisory control literature.[20]

3. **Is biased to believe and confirm:** the operation of automatic reasoning has been shown to favour the uncritical acceptance of suggestions and maintains a strong bias. Thus if a computer system suggests a target to an operator, automatic reasoning alone would make it highly likely that it would be accepted. This is known as automation bias in the supervisory literature.[21] When people seek out information to confirm a prior belief, this is confirmation bias.[22]

4. **Focuses on existing evidence and ignores absent evidence:** automatic reasoning builds a coherent explanatory story without considering any evidence or contextual information that might be missing. This is why Kahneman uses the term WYSIATI or «What you see is all there is». It facilitates the feeling of coherence that makes us confident to accept information as true whether it is or not. This is a problem if a more detailed analysis of the context of a target showed that it was not in fact legitimate. For example, an ununiformed man firing a rifle in the vicinity of an army platoon may be deemed to be a hostile target with WYSIATI. But some deliberation and a quick scan around might reveal that he had actually just killed a wolf that had taken one of his goats.

What these properties of automatic reasoning show is that in the context of supervised control of lethal targeting, things could go badly wrong. It may work well for many instances and seem OK but not when there is contradictory information of target legitimacy. Contradictory evidence could remain unseen or be disbelieved. Doubt and uncertainty will be supressed as will any notion that there is more evidence that cannot be seen.

In normal operation both automatic and deliberative processes operate smoothly

[20] J.M. Carroll and M.B. Rosson, «Paradox of the Active User», in J.M. Carroll (eds.), *Interfacing Thought: Cognitive Aspects of Human-Computer Interaction*, MIT Press, Cambridge MA 1987, pp. 80-111.

[21] K.L. Mosier and L.J. Skitka, «Human Decision Makers and Automated Decision Aids: Made for Each Other?», in M. Mouloua (eds.), *Automation and Human Performance: Theory and Applications*, Lawrence Erlbaum Associates, Inc. Mahwah NJ 1996, pp. 201-220.

[22] C.G. Lord, L. Ross and M. Lepper, «Biased Assimilation and Attitude Polarization: The Effects of Prior Theories on Subsequently Considered Evidence», *The Journal of Personality and Social Psychology*, n. 47 (1979), pp. 1231-1243.

together. The point here is that it is vitally important that deliberative reasoning is enabled in the design of supervisory control for weapons systems. Although this is also subject to error and flaws, it does as good a job as can be done with uncertainty and doubt.

If a supervisory weapons operator is distracted by another task or if they are stressed, their attentional capacity may be low. Many experimental studies have demonstrated that a small amount of interference to our attention or memory can disable the deliberative system. For example, when people are asked to do a task, such as verifying the truth or falsity of a statement, and at the same time add numbers, they will believe anything they are told to.

If a supervisory weapons operator is distracted by another task or if they are stressed, their attentional capacity may be low. So in trying to decide on the necessity or legitimacy of an attack against a target, they may not be reasoning at an acceptable level. This is one reason why, in what is known as on-the-loop control, having a single operator controlling multiple weapons systems could be disastrous. They would not be able to use their deliberative reasoning and could simply catch the downsides of automatic reasoning if there were problems or irregularities.

3. Deliberative reasoning meets supervisory control of weapons

There has been considerable confusion over the terms "autonomous" and "semi-autonomous" in discussions about weapons systems. Outside of technical robotics circles it often leads to discussions about self-governing and free will. This confusion is not helped by the many classification systems developed in an attempt to define autonomy and semi-autonomy in terms of different levels. For example the US Navy used three levels while the US Army used ten.[23] This could be very confusing for a military commander having to work with several systems at different levels.

A reframing of autonomous/semi-autonomous operation is proposed here in terms of levels of human supervisory control. This gets around the jargon and makes clear the important aspect of who is in control and how. This makes the command and control of computerised weapons systems transparent and maintains clear accountability.

An examination of scientific research on human supervisory control allows us to develop a classification consisting of five levels of control:[24]

[23] For some discussion and references see N.E. Sharkey, «Cassandra or the False Prophet of Doom: AI Robots and War», *IEEE Intelligent Systems*, n. 4, (2008), 23, pp. 14-17.

[24] This is adapted from early work on general (non-military) supervised control with 10 levels of human supervisory control: T.B. Sheridan and W. Verplank, «Human and Computer Control of Undersea Teleoperators», *Man-Machine Systems Laboratory*, Department of Mechanical Engineering, MIT, Cambridge MA 1978.

Table 2: A classification for levels of human supervisory control of weapons

1. human deliberates about a target before initiating any attack
2. program provides a list of targets and human chooses which to attack
3. program selects target and human must approve before attack
4. program selects target and human has restricted time to veto
5. program selects target and initiates attack without human involvement

For level 1 control it is critically important to adhere to strict requirements of *deliberative human control* as described in the previous section: a human commander (or operator) must have full contextual and situational awareness of the target area at the time of a specific attack and be able to perceive and react to any change or unanticipated situations that may have arisen since planning the attack. There must be active cognitive participation in the attack and sufficient time for deliberation on the nature of the target, its significance in terms of the necessity and appropriateness of attack, and likely incidental and possible accidental effects of the attack. There must also be a means for the rapid suspension or abortion of the attack.

Level 2 control could be acceptable if shown to meet the requirement of deliberative human control. A human in control of the attack would have to be in a position to assess whether an attack is necessary and appropriate, whether all (or indeed any) of the suggested alternatives are permissible objects of attack, and to select the target which may be expected to cause the least civilian harm. This requires deliberative reasoning. Without sufficient time or in a distracting environment the illegitimacy of a target could be overlooked.

A rank ordered list of targets is particularly problematic as there would be a tendency to accept the top ranked target unless sufficient time and attentional space is given for deliberative reasoning.

Level 3 is unacceptable. This type of control has been experimentally shown to create what is known as *automation bias* in which human operators come to accept computer generated solutions as correct and disregard or don't search for contradictory information.

Cummings (2004) conducted a study on an interface designed for supervision and resource allocation of in-flight GPS guided Tomahawk missile.[25] The task for operators was to decide which candidate missile from pool of 8-16 would be the correct one to redirect to a time-critical emergent target.

[25] M.L. Cummings, «Automation Bias», cit.

The impact on the speed and accuracy of decision making was tested for two different methods to redirect Tomahawks in real time.

1. The computer provided the operator with ranked recommendations including the most "optimal" missile given the situation. This is type 3 in the above classification.
2. The computer filtered all those missile that were not candidates because of physical restraints (not enough fuel etc). The operator had to decide which missile. This equates to type 2 in the above classification.

Cummings also manipulated the computer recommendations so that half the time they were correct and half the time they were wrong. The result was that Type 3 operators made significantly faster decisions overall and their accuracy was equal to the slower Type 2 operators when the computer recommendations were correct. However, when the computer recommendation were wrong, the Type 3 operators had a significantly decreased accuracy. This is known as *automation bias*: operators are prepared to accept the computer recommendations without seeking any disconfirming evidence.

Level 4 is unacceptable. It does not promote target identification and a short time to veto would reinforce automation bias and leave no room for doubt or deliberation. As the attack will take place *unless* a human intervenes, this undermines well-established presumptions under international humanitarian law that promote civilian protection.

The time pressure will result in operators falling foul of all four of the downsides of automatic reasoning described above: neglects ambiguity and suppresses doubt, infers and invents causes and intentions, is biased to believe and confirm, focuses on existing evidence and ignores absent evidence. An example of the errors caused by fast veto came in the 2004 war with Iraq when the U.S. Army's Patriot missile system engaged in fratricide, shooting down a British Tornado and an American F/A-18, killing three pilots.[26]

In the case of level 5 control there is no human involvement in the target selection and attack. As argued above, such weapons systems could not comply with international law except in very narrowly bounded circumstances.

It should be clear from the above that research is urgently needed to ensure that human supervisory interfaces make provisions to get the best level of human reasoning needed to comply with the laws of war in all circumstances.

[26] M.L. Cummings, «Automation and Accountability in Decision Support System Interface Design», *Journal of Technology Studies*, vol. 32 (2006), pp. 23-31.

4. Human supervised autonomy?

There are currently weapons systems in use that operate automatically once activated. Such SARMO (Sense and React to Military Objects) weapon systems intercept high-speed inanimate objects such as incoming missiles, artillery shells and mortar grenades automatically. Examples include C-RAM, Phalanx, NBS Mantis and Iron Dome. These systems complete their detection, evaluation and response process within a matter of seconds and thus render it extremely difficult for human operators to exercise meaningful supervisory control once they have been activated other than deciding when to switch them off.

These are precursors to fully autonomous weapons according to Human Rights Watch.[27] Others have separated them from fully autonomous weapons by calling them automated or automatic systems. The International Committee of the Red Cross propose that: «An automated weapon or weapons system is one that is able to function in a self-contained and independent manner although its employment may initially be deployed or directed by a human operator».[28] The US department of defence attempted to bound the scope of these weapons by suggesting that: «[…] The automatic system is not able to initially define the path according to some given goal or to choose the goal that is dictating its path».[29]

There are a number of common features for SARMO weapons[30] that are necessary although not sufficient to keep them within legal bounds:

- fully pre-programmed to automatically perform a small set of defined actions repeatedly and independently of external influence or control;
- used in highly structured and predictable environments that are relatively uncluttered with very low risk of civilian harm;
- fixed base – although these are used on manned naval vessels, they are fixed base in the same sense as a robot arm on a ship would be;
- switched on after detection of a specific threat;
- unable to dynamically initiate a new targeting goal or change mode of operation once activated;

[27] Human Rights Watch and Harvard Law Clinic, «Losing Our Humanity: the Case Against Killer Robots», *Human Rights Watch report*, November 2012.
[28] ICRC, *International Humanitarian Law and the challenges of contemporary armed conflicts*, Geneva, 28 November to 1 December 2011, p. 39.
[29] US Department of Defense, *Unmanned Systems Integrated Roadmap, FY 2013-2038*, 2013, p. 66.
[30] Fire and forget weapons such as radiation detection loitering munitions and heat seeking missiles are not included here and require a separate discussion.

- have constant vigilant human evaluation and monitoring for rapid shutdown in cases of targeting errors, change of situation or change in status of targets;
- the output and behaviour of the system is predictable;
- only used defensively against direct attacks by military objects.

The US Department of Defense call these human supervised autonomous weapons: «Human-supervised autonomous weapon systems may be used to select and engage targets, with the exception of selecting humans as targets, for local defense to intercept attempted time-critical or saturation attacks for: (a) Static defense of manned installations; (b) Onboard defense of manned platforms».[31]

From the perspective of the human supervisory control framework proposed here, it is the human decision of when to use the weapon that is key to the legality of SARMO weapons systems. It is essential for making such decisions that precautionary measures have been taken about the target significance – its necessity and appropriateness, and likely incidental and possible accidental effects of the attack i.e., precautionary measures have been taken.[32] It is also essential that vigilance is maintained during operation of the weapons systems and that there is a means for rapidly deactivating the weapons if it becomes apparent that the objective is not a military one or that the attack may be expected to cause incidental loss of civilian life.[33]

One concern arises from the justification used by the UK Ministry of Defence for SARMO weapons.[34] It leads into very dangerous legal territory: «The potential damage caused by not using C-RAM in its automatic mode *justifies the level of any anticipated collateral damage*» [italics mine]. This omits precaution, proportionality and necessity and is unacceptable under International Humanitarian Law.[35] Such justifications could lead us into the incautious use of unsupervised weapons systems that could cause disproportionate harm to civilian populations and objects.

The precautionary principle cannot be overstressed for the use of so called "supervised autonomy". We must be wary and vigilant about the mission creep. As the UK Ministry of Defence point out: «The role of the human in the loop has, before now, been a legal

[31] US Department of Defense, *Autonomy in Weapon Systems*, cit., p. 7.
[32] As specified Article 57 of additional protocol 1 to the Geneva Convention 1977. Downloaded from: http://bitly.com/1hJF4GC Last accessed March 5 2014.
[33] Ivi, see for a full account the points 2: (a), iii, and (b).
[34] [34] UK Ministry of Defence, *Development, Concepts and Doctrine Centre, The UK Approach to Unmanned Aircraft Systems*, Joint Doctrine Note, March 30 2011.
[35] Thanks to Dr. Nils Melzer from the Geneva Centre for Security Policy, personal communication, for pointing this out.

requirement which we now see being eroded […]».[36] But it is essential that we avoid such erosion and lock down human supervisory control as a legal principle of human control.

As an example of erosion, we can look at a possible extension to the German NBS Mantis system. The Mantis uses the speed and trajectory of an incoming mortar grenade to calculate where to fire «a cone shaped metal cloud» of projectiles into its path. The response time for the system to detect and fire is expected to be approximately 4.5 second. However, according to the manufacturer's specification: «MANTIS' control system is also capable of tracking the location of the assailants along with the flight path and point of impact».

The use of the word "assailants" is problematic here. All that can be detected from the speed and trajectory of the incoming munitions is from where they were fired. It should not be assumed that the assailants are present at the location. This information can certainly be useful for a commander to locate and assess whether or not there are legitimate targets at the location, their necessity and appropriateness and whether or not an attack on them would be proportionate. It is possible, for example, that the mortar was set in a civilian urban area and fired by remote control.

It is not been suggested that the information from the Mantis would be used to automatically attack the area from where the munitions were launched, but it is an obvious extension that needs to be ruled out for such weapons systems. It would mean that they had stepped outside of SARMO functionality.

5. Conclusions

A number of high-tech states are moving forward with the development of fully autonomous platforms or robots with plans for them to carry weapons. A key question probed here was, how will the proposed onboard weapons systems be controlled.

The current state of Automatic Target Recognition was evaluated and found wanting on a number of counts. The bottom line was that fully autonomous weapons, those that once activated would select targets and attack them without further human intervention, could not be used in a way that could be guaranteed to predictably comply with International Law.

There is general agreement on the inadequacy of Automatic Target Recognition and some states such as the US and UK have made it clear that, at least for the time being, computerized weapons systems will always be under human control. What has not been made clear however, is what type of human control will be employed and how meaningful it will be. Thus the main aim of this article was to pull apart and examine the minimum

[36] UK Ministry of Defence, *Development, Concepts*, cit.

necessary condition for the notion of meaningful control.

It was noted that both humans and computer systems have their strengths and weaknesses and that the aim of designing effective supervisory control systems for weapons control is to exploit the strengths of both. In this way it may be possible to gain better legal compliance than with either humans or computer systems operating on their own.

It is imperative that humans are not being asked to perform in a mode of operation that is not well suited to human psychology. This needs to be part of the equation of ensuring efficient and meaningful human supervisory control of weapons. If the interface with the computing system is not right for humans, they will fail. But that should not be taken as a good reason for saying that machines could do the task better. It is simply a good reason for saying that we need a better-designed interface.

Much of the article focused on the nature of human reasoning, the strengths and weakness of different type of reasoning processes and how they would impact on the supervisory control of weapons systems. This was followed by a reframing of autonomous operation in terms of levels of supervisory control. A new five-level human supervisory control framework was proposed and discussed.

The legitimacy of, what has been called, supervised autonomous weapons systems was discussed in terms of their properties and the minimum legal requirements for their operation. Caution was urged about how such systems could be extended in a way that allowed them to stray beyond the limits of appropriate targeting supervision. We must not allow the accelerating pace of warfare to dictate that we should use computerised weapons that are not meaningfully controlled by human operators. It is imperative that we develop a principle of human control founded on human reasoning processes to provide clear guidelines for state weapons reviews.