ICRAC

International Committee for Robot Arms Control

Guidelines for the human control of weapons systems

April 2018

WORKING PAPER FOR CCW GGE

Guidelines for the human control of weapons systems

Authored by Noel Sharkey, chair of ICRAC[1]

Since 2014, high contracting parties to the CCW have expressed interest and concern about the meaningful human control of weapons systems. There is an extensive scientific and engineering literature on the dynamics of human-machine interaction and human supervisory control of machinery. A short guide is presented here consisting of two parts. Part 1 is a simple primer on the psychology of human reasoning. Part 2 outlines different levels for the control of weapons systems, adapted from human-machine interaction research, and discusses them in terms of the properties of human reasoning. This outlines which of the levels can ensure the legality of human control of weapons systems and guarantee that precautionary measures are taken to assess the significance of potential targets, their necessity and appropriateness, as well as the likely incidental and possible accidental effects of the attack.

1. A short primer on human reasoning for the control of weapons

A well-established distinction in human psychology, backed by over 100 years of substantial research, divides human reasoning into two types: (i) fast *automatic* processes needed for routine and/or well tasks like riding a bicycle or playing tennis and (ii) slower *deliberativ*e processes needed for thoughtful reasoning such as making a diplomatic decision.

The drawback of deliberative reasoning is that it requires attention and memory resources and so it can easily be disrupted by anything like stress, or being pressured into making a quick decision.

Automatic processes kick in first, but we can override them if we are operating in novel circumstances or performing tasks that require active control or attention. Automatic processes are essential to our normal functioning, but they have a number of liabilities when it comes to making important decisions such as those required to determine the legitimacy of a target.

Four of the known properties of automatic reasoning[2] illustrate why it is it problematic for the supervisory control of weapons.

- **neglects ambiguity and suppresses doubt**. Automatic processes jump to conclusions. An unambiguous answer pops up instantly without question. There is no search for alternative interpretations or uncertainty. If something looks like it might be a legitimate target, in ambiguous circumstances, automatic reasoning will be certain that it is legitimate.
- **infers and invents causes and intentions.** Automatic reasoning rapidly invents coherent causal stories by linking fragments of available information. Events that include people are automatically attributed with intentions that fit a causal story. For example, people loading muckrakes onto a truck could initiate a causal story that they were loading rifles. This is called *assimilation bias* in the human supervisory control literature.[3]
- **is biased to believe and confirm.** Automatic reasoning favours uncritical acceptance of suggestions and maintains a strong bias. If a computer suggests a target to an operator,

[1] Special thanks to Lucy Suchman, Frank Sauer and Amanda Sharkey and members of ICRAC for helpful comments

[2] D. Kahneman 2011:, Thinking, Fast and Slow, Penguin Books. He refers to the two processes as System 1 and System 2, These are exactly the same as the terms automatic and deliberative used here for clarity and consistency.

[3] 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, 1987), 80–111.

automatic reasoning alone would make it highly likely to be accepted. This is *automation bias*.[4] *Confirmation bias*[5] selects information that confirms a prior belief.

- **focuses on existing evidence and ignores absent evidence.** Automatic reasoning builds coherent explanatory stories without consideration of evidence or contextual information that might be missing. What You See Is All There Is (WYSIATI)[6]. It facilitates the feeling of coherence that makes us confident to accept information as true. For example, a man firing a rifle may be deemed to be a hostile target with WYSIATI when a quick look around might reveal that he is shooting a wolf hunting his goats.

2. Levels of human control and how they impact on human decision-making

We can look at levels of human control for weapons systems by adapting research from the human supervisory control literature as shown in Table 1.[7]

A classification for levels of human supervisory control of weapons

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

Level 1 control is the ideal. A human commander (or operator) has full contextual and situational awareness of the target area at the time of a specific attack and is able to perceive and react to any change or unanticipated situations that may have arisen since planning the attack. There is 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, and likely incidental and possible accidental effects. There must also be a means for the rapid suspension or abortion of the attack.

Level 2 control could be acceptable if it is shown to meet the requirement of deliberating on potential targets. The human operator or commander should deliberatively assess necessity and appropriateness and whether any of the suggested alternatives are permissible objects of attack. 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 automation bias could create a tendency to accept the top ranked target unless sufficient time and attentional space is given for deliberative reasoning.

[4] K.L. Mosier and L.J. Skitka 1996: Human decision makers and automated decision aids: made for each other?, in: Mouloua, M. (Eds.): Automation and Human Performance: Theory and Applications, Lawrence Erlbaum Associates, 201–220.

[5] C.G. Lord, L. Ross and M. Lepper 1979: 'Biased assimilation and attitude polarization: the effects of prior theories on subsequently considered evidence', Journal of Personality and Social Psychology, 47, 1231–1243.

[6] Kaheneman ibid.

[7] For a more in-depth understanding of these analyses and references see N. Sharkey 2016: Staying in the Loop. Human Supervisory Control of Weapons, in: Bhuta, Nehal et al. (Eds.): Autonomous Weapons Systems. Law, Ethics, Policy. Cambridge University Press, 23-38.

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 trust computer generated solutions as correct and disregard or don't search for contradictory information. Cummings experimented with automation bias in a study on an interface designed for supervision and resource allocation of in-flight GPS guided Tomahawk missiles.[8] She found that when the computer recommendations were wrong, operators using Level 3 control had a significantly decreased accuracy.

Level 4 is unacceptable because it does not promote target validation 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 neglecting ambiguity and suppressing doubt, inferring and inventing causes and intentions, being biased to believe and confirm, focusing on existing evidence and ignoring absent but needed evidence. An example of the errors caused by demands of fast veto was in the 2004 Iraq war when the U.S. Army's Patriot missile system shot down a British Tornado and an American F/A-18, killing three pilots.

Level 5 control is unacceptable as it describes weapons that are autonomous in the critical functions of target selection and the application of violent force.

It should be clear from the above that there are lessons to be drawn both from the psychology of human reasoning and from the literature on human-machine interaction. An understanding of this research is urgently needed to ensure that human-machine interaction is designed to get the best level of human control needed to comply with the international law in all circumstances.

Conclusion: Necessary conditions for meaningful human control of weapons.

A commander or operator should

1. have full contextual and situational awareness of the target area at the time of initiating a specific attack;
2. be able to perceive and react to any change or unanticipated situations that may have arisen since planning the attack, such as changes in the legitimacy of the targets;
3. have active cognitive participation in the attack;
4. have sufficient time for deliberation on the nature of targets, their significance in terms of the necessity and appropriateness of an attack and the likely incidental and possible accidental effects of the attack and…
5. have a means for the rapid suspension or abortion of the attack.

[8] M.L. Cummings 2006: Automation and Accountability in Decision Support System Interface Design, in: Journal of Technology Studies 32: 1, 23–31.