HUMAN CONTROL OF WEAPONS SYSTEMS

Noel Sharkey
International Committee for Robot Arms Control

Since 2014, high contracting parties to the Convention on Conventional Weapons (CCW) have expressed interest and concern about the meaningful human control of weapons systems. Different states use different terms from 'appropriate levels of human control' and 'person in the loop' to the 'wider loop'. There is also the notion of force multiplication with one or two people operating a swarm of weapons systems. Yes, these are all forms of human control, but the important question is, what kind of human control is necessary to 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?

This chapter shares guidelines that have been designed to provide campaigners with tools to assess whether proposed methods of control are meaningful or not. It delivers a plain English guide to lessons learned from 30 years of scientific research on the human supervisory control of machinery.

Part 1 is a primer on the study of human reasoning. It briefly explains the types of biases that result in operators making bad decisions and it explains the kind of reasoning needed for meaningful human control?

Part 2 puts the primer on human reasoning to work to show the types of human control that are unacceptable for making targeting decisions.

1. SHORT PRIMER ON HUMAN REASONING FOR THE CONTROL OF WEAPONS

A well-established distinction in human psychology divides human reasoning into two types:

i. fast *automatic* processes that are needed to carry out routine everyday tasks like riding a bicycle, avoiding traffic or playing a sport. This is vital when we need to react quickly or carry out a task without engaging our conscious thought.

ii. slower *deliberative* processes that are needed for thoughtful reasoning. This is important for making important judgements such as diplomatic, medical or judicial decisions and, hopefully, even decisions about getting married or divorced.

One drawback of deliberative reasoning is that it can be fragile. It requires attention and memory resources and so it can easily be disrupted by stress or being pressured into making very quick decisions.

Automatic reasoning is essential to our normal daily functioning, but it has a number of liabilities when it comes to making important decisions such as those required to determine the legitimacy of a target.

"Automatic reasoning jumps to conclusions."

Four of the known properties of automatic reasoning illustrate why it creates problems for the control of weapons. Automatic reasoning:

- **neglects ambiguity and suppresses doubt.** Automatic reasoning jumps 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, in the context of an armed conflict people loading rakes onto a truck could initiate a causal story that they were loading rifles. This is called *assimilation bias* in the human supervisory control literature.

- **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, automatic reasoning alone would

make it highly likely to be accepted. This is *automation bias. Confirmation bias* 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). 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.

It should be clear that each of these features of automatic reasoning would lead to serious humanitarian errors. When people talk about various types of human in the loop control systems or controlling a swarm, we need to look carefully to find out if they trap the operator in the error-prone properties of automatic reasoning.

2. LEVELS OF HUMAN CONTROL AND HOW THEY IMPACT ON HUMAN DECISION-MAKING

Now that we have looked at some of the relevant properties of human reasoning, we can see what that tells us about the control of weapons. In the science world, different way to control machinery are discussed in term of levels. Level 1 would be the best and level 5 would be unacceptable.

In Table 1, the machinery levels have been adapted to describe levels of controlling weapons. These should not be considered to be definitive or absolute. The levels are intended as thought tools to help you to work out whether some new human control method stacks up.

A classification for levels of human 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 and confirmation bias could take hold.

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.

Level 3 is unacceptable. This type of control has been shown to create as automation bias in which human operators come to trust computer generated solutions as correct and disregard or don't search for contradictory information. Studies on automation bias in the supervision of Tomahawk missiles found that when the computer recommendations were wrong, operators using Level 3 control had tended to treat them as correct. Level 1 operators were a little slower when things went well but performed well when computer recommendations went wrong.

Level 4 is unacceptable because it does not promote target validation and a short time to veto and attack 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.

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.

IN SUMMARY

It should be clear from the above that there are many types of control that would not fulfil the conditions of **Level 1** control. You should be in a position now to ask questions about any method of control and find out how it fits in the Levels shown in Table 1. The biases and problems with automatic reasoning described in Part 1 will help you to assign the correct level. It might be between two different levels or it might need an entirely different level. Working in this way should assist in determining risks to International Humanitarian Law.