Ethics and the killer robot

Murray Sinclair & Carys Siemieniuch

Recently, through no fault of ours, one of us became involved in a formal debate on the following motion: 'This house believes that the engineers who design lethal autonomous and semi-autonomous systems must accept ultimate responsibility for the ethical behaviour of those systems.' In the final vote, this motion was accepted. However, it seemed that we could also have inserted 'ergonomists' in place of 'engineers'. For this reason, we pass on some of the thinking. First, let's set the scene, and then discuss two scenarios.

What do we mean by 'lethal'? The debate was about military robots deployed in the battlespace with weapons and programmed to kill, given the right circumstances. However, the discussion could also apply to 'peaceful' robots that can kill inadvertently.

What do we mean by 'autonomy'? The essence of autonomy is that the robot has both the capability to make its own decisions, and has access to resources to enable to carry out its decisions. There are degrees of autonomy:

1. The computer offers no assistance; humans must do it all.
2. The computer offers a complete set of action alternatives,
3. and narrows the selection down to a few,
4. or suggests one,
5. and executes that suggestion if the human approves,
6. or allows the human a restricted time to veto before automatic execution,
7. or executes automatically, then necessarily informs the human,
8. or informs the human after execution only if it is asked,
9. or informs the human after execution if the computer decides to do so.
10. The computer decides everything and acts autonomously, ignoring the human

Given that these robots will be in the armed services, they will be within a hierarchy of control. Hence, we need two other dimensions. Firstly, what autonomy is allocated to the robot (e.g. it can kill over there, but not here, and over there it must follow the Rules of Engagement). Secondly, there is the amount of time for which the robot is allowed to be autonomous (the less the time, the less the damage the robot can do). Finally, it is likely that a robot's allowable or acceptable autonomy will be affected by the weapons it carries. If it is armed with a pea-shooter, nobody will worry much about its level of autonomy. If it has a nuclear weapon on board, we might be very concerned.

What is ethical behaviour? In essence, in a war situation the basis of ethical behaviour is provided by the Law of Armed Conflict (LOAC) in tandem with the Rules of Engagement that apply to a particular conflict. LOAC does not specify that decision making in the battlespace must be carried out by humans; therefore, if decisions leading to intentional death are made, it does not seem to matter whether this process is undertaken by humans or machines. LOAC is emphatic about the requirement for military necessity, the proper discrimination of targets and non-targets, proportionality in the use of firepower, and the embodiment of the principles of humanity in decisions.

Who has 'ultimate' responsibility? Note that 'ultimate' does not imply 'total'; merely that you are fundamentally implicated. Your responsibility could become total if the robot that you have designed has the capability to understand unethical or illegal commands, and is capable of refusing to enact them. But ou will hardly be held responsible if the robot has been given imperfect information.

Learning! This is a big problem. Autonomous systems in general, not just robots, need to be given the capability to learn, so that they can learn from their mistakes and can become more proficient in the range of environments in which they function. So how do we ensure that our robot learns nothing that is unethical? Are we responsible if it does, since we have given it its learning capability? Let's consider two examples.

Example 1. Energetically Autonomous Tactical Robot (EATR), a DARPA-funded prototype developed in the USA. It's "able to perform long-range, long-endurance missions without the need for manual or conventional re-fueling, which would otherwise preclude the ability of the robot to perform such missions". That's because it's designed to live off the land, off biomass, like a goat or a sheep. It gathers food by means

of a gripper and a chainsaw and pops it in a hopper on top. It then produces electricity to power itself. So in principle it could be sent off into lower Helmand, armed, to patrol the border with Pakistan. They say that it needs up to 70 kg of dry mass per day for full roving functionality. Compare this to a sheep, which needs up to 2kg dry mass per day. Sheep spend 5-10 hours per day feeding, depending on the quality of the food so the EATR could spend most of its time in Helmand roaming for food. And this raises some ethical issues. Suppose it comes across an old fence. LOAC says that civilian artefacts shall be left alone. Suppose the robot finds a Taliban corpse out there, after an airstrike. Assuming the body is fully clothed, behold! dry food, albeit wrapped round a corpse. What happens next? LOAC is very specific about not desecrating the dead, so whether the robot does anything may depend on its powers of discrimination. In such a scenario, it's difficult to see that the robot's commander would be held responsible for what the robot did, unethically, more likely it's those who designed and maintained it.

Example 2. US Department of Homeland Security (DHS) (adapted from an example in Wallach & Allen). It's morning in America, on May 2nd, 2015 and it's the first really hot and humid day of the year. The electricity producers have decided to start up some old, coal-fired plants to meet the expected demand for air conditioning. It's 09.15 and one of the plants near New York has a turbine problem, followed by a small explosion and the plant goes into emergency shut-down, resulting in a few areas having a power black-out. The DHS computers receive this information (explosion; shut-down; blackouts) and deduce there's a possibility it's due to terrorist action. Then, due to another unrelated incident - a small aircraft hitting an electricity pylon in Illinois - the computers decide to raise the security threat level. At Reagan Airport near Washington DC the security computers, scanning passengers' biometric data, tighten their criteria in the light of the raised security level. They detect that five passengers going to London and on to Karachi should be re-inspected, but the flight is boarding, so armed security guards rush along and grab them. There's a scuffle and a gun is fired. It's 09.20 when the DHS computers learn of this and upgrade the threat level again. Meanwhile in Texas, there are autonomous weapons, connected to the DHS computers, mounted back from a fence running along the border to deal with drug-smuggling. These have 'stop and desist' signs connected to them. In an empty part of the state, one of these weapons detects a Hummer coming out of the desert and lights up its 'stop' sign. The Hummer stops, and this being Texas, someone leans out and shoots the sign. Because of the high threat level the autonomous weapon immediately fires in response, and the Hummer is destroyed. It's 09.25. While LOAC itself does not apply to this situation, its principles are still valid. The Hummer victims died as a result of an over-reaction by a computerised system-of-systems, in disregard of LOAC principles. Of course, there will be humans within this system of systems, nominally in charge, and able to over-rule the decisions. But they will only do this with secondary, confirming evidence, because nobody will want to risk saying "It's all OK", and then find they really are dealing with a terrorist incident. This takes time, and the whole scenario could be over before they have confirmation. In this system of systems, it is likely that the different components were developed by different organisations and brought into interoperation at different times by different people, not necessarily aware of what other changes are or have been happening in the system-of-systems. Hence, even though you may not be working on lethal autonomous systems, if they can be connected to your work, you may have to consider the ethical issues of what you are creating.

So where does this leave us?

♦ Autonomous systems necessarily must embody ethics in their design, development and operation. This is particularly the case for intentionally-lethal systems, and is likely to be true for those that could kill or maim inadvertently. In a system-of-systems context, ethical considerations will be spread across the whole network of systems.

♦ We don't yet know how to embody ethics in an autonomous system, especially if they are given the capability to learn. Given an ethics disaster, it's likely that both the controllers of these autonomous systems and their developers will be queried in a court of inquiry as to their responsibility for the disaster. To the extent that ergonomists are involved in both of these domains, we might be present, too.

♦ We might have identified another interface between humans and machines; the ethics interface. If we have, then we need to understand it. This is a next generation issue and this means you! ❖