Aiding designers, operators & regulators to deal with legal & ethical considerations in the design & use of lethal autonomous systems

M J de C Henshaw[1], C. E. Siemieniuch[1], M. A. Sinclair[2] ,
[1]*Department of Electronic and Electrical Engineering,*
[2]*Centre for Innovative and Collaborative Engineering, Department of Civil and Building Engineering*
Loughborough University, Leicestershire, LE11 3TU, UK
c.e.siemieniuch@lboro.ac.uk

Abstract

The focus of this paper is how to design legal and ethical behaviour into Semi/Autonomous Systems (S/AS) that work within a System-of-Systems (SoS) context, in either a military or civilian environment. Its intention is to explore the problem domain, question the comprehensiveness of current work in this area and postulate a series of issues that the authors feel need to be addressed if serious progress is to be made in an area of considerable interest to governments, manufacturers and users of theses systems and indeed the general public. It should be read as a discussion document.

1. Background rationale

Consider the following extracted from [1, 2] "Hakimullah Mehsudis probably dead …. was the target for a missile fired last month from an unmanned aircraft hovering over the Afghan-Pakistani border - but launched by an operator in the US. …Mehsud was ….. part of a suicide mission on December 30 at Khost, just across the border in Afghanistan, which killed seven CIA agents ……. However the Americans confirmed the breadth of a new horizon in modern warfare - launching 15 clinical drone attacks in which more than 100 people died along the border, as Washington's electronic eyes and guns sought out Mehsud and his Taliban and al-Qaeda allies. …….. Consider: for the first time ever, a civilian intelligence agency is manipulating robots from halfway around the world in a program of extrajudicial executions in a country with which Washington is not at war."

"According to the residents of Datta Khel, a town in Pakistan's North Waziristan, three missiles streaked out of Afghanistan's Pakitka Province and slammed into a Madrassa, or Islamic school, this past June. When the smoke cleared, the *Asia Times* reported, 30 people were dead. The killers were robots, General Atomics MQ-1 Predators. The AGM-114 Hellfire missiles they used in the attack were directed from a base deep in the southern Nevada desert

Automatic, semi-autonomous and autonomous agents/systems (S/AS) are increasingly used for a range of warfare /peacekeeping functions (the 3 'Ds'). Several S/AS already use lethal force (eg Cruise & Patriot missiles, Predator). However, as levels of autonomy increase it will be necessary to differentiate the use of S/AS: are they merely 'weapons', an extension of the war-fighter – currently the Law of Armed Conflict (LOAC) and the Rules of Engagement (RoE) apply; or potential autonomous agents which reserve the right to make their own local decisions regarding the application of lethal force in the field without requiring human consent. There are a range of legal, moral and ethical issues that need to be addressed in both of these contexts but particularly the latter. Clausewitz [3] put it well; "In war the will is directed at an animate object that reacts." However this 'will' needs to be controlled and managed to avoid incidents such Seboka training exercise in S Africa in 2007, which left 9 soldiers dead and 14 seriously injured [4].

These concerns apply equally to 'peaceful' S/AS such as 'industrial' robots that can kill inadvertently. Some 4 per cent of US factories have "major robotics accidents" every year: a man having molten aluminium poured over him; a woman picked up and placed on a conveyor belt to be smashed into the shape of a car; even former Japanese Prime Minister Junichiro Koizumi was nearly killed a few years ago after a robot attacked him on a tour of a factory [5]. In addition there is some cross-over of issues to quite different domains such as automated financial trading where legal, ethical, regulatory and moral dimensions pertain. The decisions may be financial; the consequences will be human.

2. Current work in this area

This issue has been recognised by many authors. So far, they tend to fall into two camps. The first of these is the philosophical/legal camp, which discusses the ethics issues in great detail and in general reaches the conclusion that ethical autonomous systems are unlikely to eventuate [5-8]& [9].

978-0-7695-4175-4/10 $26.00 © 2010 Crown Copyright
DOI 10.1109/EST.2010.35

The other camp is the engineering camp, which is gradually incrementing its development of autonomy in systems but still seems to be a long way from introducing ethical considerations (apart from those concerned with safety) [10, 11]12]. It is the gap between these camps that needs to be explored - where scenarios of use can be generated and explored for their ethical, engineering implications.

These kinds of scenarios have already generated some findings. For example, it is not enough to make the weapon-carrying robot able to assess its plans for ethical issues; for reasons explained later the weapons must bring their own ethics with them as well. Consequently we have to consider an 'ethics interface'. Secondly, this interface will have to be open to human oversight. Thirdly, as systems-of-systems grow and necessarily become more autonomous in their decision-making, to the extent that they can issue commands to armed robots, they too must be capable of ethical reasoning; it cannot be left to humans and armed robots and not the interconnecting network with its information-processing applications. Furthermore, since the Law of Armed Conflict (LOAC) and the Rules of Engagement (RoE) require considerable interpretation in order to assess a developing situation, and no one is clear how this should be done, one can foresee that different ethical agents within the system of systems may get into the same arguments about what can and can't be actioned as at present among human agents. Finally, and perhaps more importantly for society, it is also evident that culture and politics cannot be separated from ethical behaviour by autonomous systems.

Three elements are important here:

- **Ethics** comprise the standards or codes of behavior expected by the group to which the individual belongs. There may be several of these codes; e.g. national ethics, social ethics, company ethics, professional ethics, or family ethics. Ethics provide the norms of behaviour that society expects. Because different societies may have different ethical codes, in a 'Coalition of the Willing' the same situation may engender different behaviours. Laws and Regulations in society can be considered to be formalisations of ethics.
- **Morals** define personal character, while ethics stress a social system in which those morals are applied. When morals and ethics are co-aligned, life is easy. When they are not, problems arise (e.g. anti-social behaviour, criminal acts). A person's moral code is fairly stable, and slow to change; the ethics he or she practices are other-dependent, and can change relatively quickly – e.g. by Acts of Parliament.
- Culture (Cultural Attributes) is the unconscious acquisition of a common set of values, preferences, attitudes, beliefs and rules that influence individual and group emotions, and their behaviours toward individuals, group, the environment and other artefacts:

3. Core Problem Focus

The high level problem is how to design legal and ethical behaviour into Semi/Autonomous Systems (S/AS) that work within a System-of-Systems (SoS) context, in either a military or civilian environment. Current practice tends towards controlling system use by Regulation (Law of Armed Combat, Rules of Engagement, Health and Safety, Certification, Operational Procedures etc), which impacts at the operational end of the Engineering Life Cycle but which needs be addressed at the earlier phases. To ensure Fitness for Purpose the question becomes how to provide engineers/designers with tools and techniques that allow them to translate the 'legal and ethical use' concerns embedded in the various regulatory instruments into clear requirements, design specifications and architectural considerations at the concept, bid and other design phases in order to limit the potential for undesirable or lethal behaviour at the operational phase.

The second, inter-related part is regulation. In general, regulatory frameworks are there to condition and constrain human behaviour and decision making. As well as written regulations they frequently encompass authorities (individuals, committees, boards etc) that provide interpretation and judgement, either during design and development (eg certification) or during use (eg doctrine/investigation/adjudication). How such frameworks and principles should evolve with increasing system autonomy is at the heart of the required research. This issue is exacerbated by the paradigm shift towards product service and support in many engineering domains.

4. Issues that need to be addressed

Based on the discussion above and the outline of the core problem to be addressed, the authors have identified a number of issues that should be considered so that effective progress can be made in this area. Since this paper is intended to provoke discussion rather than provide answers or proffer solutions, these issues are listed below. However the authors make no claim that this is a complete set – rather they should be seen as a starting point for further research and exploration.

1 The first part of the approach should be to define the problem result that is required, then construe what needs to be done. Initially, the problem is to get robots to behave ethically, in relation to LOAC, local ROE, and (for the UK) the Convention on Human Rights. In other words, using Moor's words [13], we don't have to go fully to 'explicit ethical systems'; it

would be enough to go for ‘implicit ethical systems’ first. The latter are systems that behave ethically, but can’t reason ethically; their ethics come from the hardware design of the system (including safety), with ethics arriving from the outside via current instructions.

2 The second part of the problem is to chip off bits of the problem (with care, and recording the logic) until we can see how to get a solution, then incrementally grow this solution until it can cover enough cases to be usable.

3 The third part of the problem is that we are not concerned only with the individual robot acting ethically, it is the swarm that should be the focus. Note the implications of this; we have a distributed ethics network over platforms that may have to collaborate.

4 An important aspect of the problem (which maybe will make it simpler) is to embrace the fundamental issue of time. A decision is not often a one-off; it is usually one within a sequence. Hence, if we can ensure that the decision-making robot preferentially makes decisions which can be undone or retrieved or repaired, we may simplify the ethical issues. Furthermore, we have a context in which the decisions can be made, and in which particular sets of ethics will be more appropriate, dependent to some extent on the ethics the system has been applying for the last decision, and to a lesser extent those that went before. . This comment makes use of the notion that given a current decision in a given situation, there is a limited number of alternative decisions that are logically available (unless a discontinuity suddenly appears in the evolving situation). Again, this may reduce the complexity of our problem.

Furthermore, we can considerably simplify the issues by bringing in the concept of regulation. It fits for simplifying human behaviour, and will certainly reduce what we need to consider - see points (4), (7), (10) and (11). Since it is highly unlikely that the military (or Police) will allow robots to be off the leash until some time in the future, we may be able to reduce the range of ethical behaviour to be considered

5 Note that (4) above implies that armed and dangerous robots will kill only if there is ‘no other alternative’. This is likely to be a desirable characteristic (the threat of death, rather than instant death), which differentiates this approach from that of others (e.g. [11]). In a world of asymmetric warfare, this might be a better approach. Also note that this doesn’t mean the robot has to be reactive; it can carry out aggressive search, discriminate, and confirm, perhaps within a swarm of robots, but once combatants have been identified, it means the robot doesn’t have to assault them immediately.

6 Note also that (4) and (5) imply that it is not just the ethics of the robot which are of concern; it includes the ethics of the weapon, including how much damage it can do. In other words, weapons must come with ethics attached, that the robot can utilise. Nobody is much bothered about the ethics of a robot equipped with a water-pistol, but if it carries a tactical nuclear weapon, then everybody is concerned. Similarly for the swarm; if they represent different lethal capabilities, we need to be careful concerning their interoperation - lo, we have a new interface to describe; the ethics interface.

In addition, once you take integrity of the system into account, the weapon needs to interrogate the delivery platform to make sure it is ours, and not in the hands of the Bad Guys. This becomes more and more important as the weapon’s bang becomes bigger, so we are probably in the domain of use-once passwords and encryption. But note also an additional confirmation route might be for the weapon to ask the robot (and vice versa) ‘Where have you been, recently?’. Hence, architecturally, we may need to capture GPS co-ordinates. It follows that, architecturally, both the robot and the weapon should have similar ethics systems. By adopting this approach we may also simplify the issues of fitting several different weapons on a single robot platform.

7 Discussing the notion of the ‘swarm’ suggests there will need to be some higher level of co-ordination. If we keep this whole problem as simple as possible, initially, then this level of co-ordination/ control will be exercised by humans. A scenario would be that a soldier is in command of a section of robots, and is tasked with clearing an area. A loose analogy would be a shepherd and sheepdogs, rounding up a flock on the high hills - except that we would want our robots to have a bit more ethics in their behaviour. Does a dog have ethics? Unknown, but it does have behavioural patterns, and it can be trained not to damage or kill the sheep (implicit ethical behaviour). Note also that what the dog does is sequential, with some re-ordering and repetition of behaviour when the sheep don’t do what is expected.

8 A particular worry is that autonomous, mobile robots have to be able to learn. Hence, how do we ensure that a robot does not ‘learn’ to ignore ethics? It seems that we need some way either to exclude any onboard ethics module from learning, or to download ethics to the robot each day, or to upload what the robot has learnt about ethics during the day, check it for good/bad learning, remove the bad, and download the good to all associated robots.

Note the implication of this. Ethical behaviour is partly determined outside the robot. Again, this might simplify the problem, but it creates another; what is the interface that allows a human to check and modify what the robot has learnt?

9 So far, we have excluded the environmental recognition/ discrimination problem ("what is going on?") and treated it as outside the boundary of the ethics problem. But this is a fuzzy boundary; given we are concentrating on time-based ethics, it is likely that a part of the recognition problem (is this traditionally-dressed woman coming towards us a bomber?) may involve ethics (if she is not a bomber, would she be doing this?). Let us label this as a 'much later' problem, for attention once we have the ethics apparatus to contribute to this.

10 One further issue; how does the robot distinguish between peace, war games and war? In peace, it should not be armed, except when on sentry duty at sensitive sites. In this case, presumably the robot should consider itself at war, within its immediate environment. War games extend the need for different rules for lethal action.. The robot identifies 'Bad Guys', follows its rules, and shoots. The weapon acknowledges it has shot, but the Bad Guys are evidently still alive. So, what does the robot do? Shoot again? Significant security issues manifest themselves at this point.

5. What would help designers?

Figure 1 below indicates the current state of armed robotics – the MAARS robot from Boston Dynamics, with multiple weapons mounted.

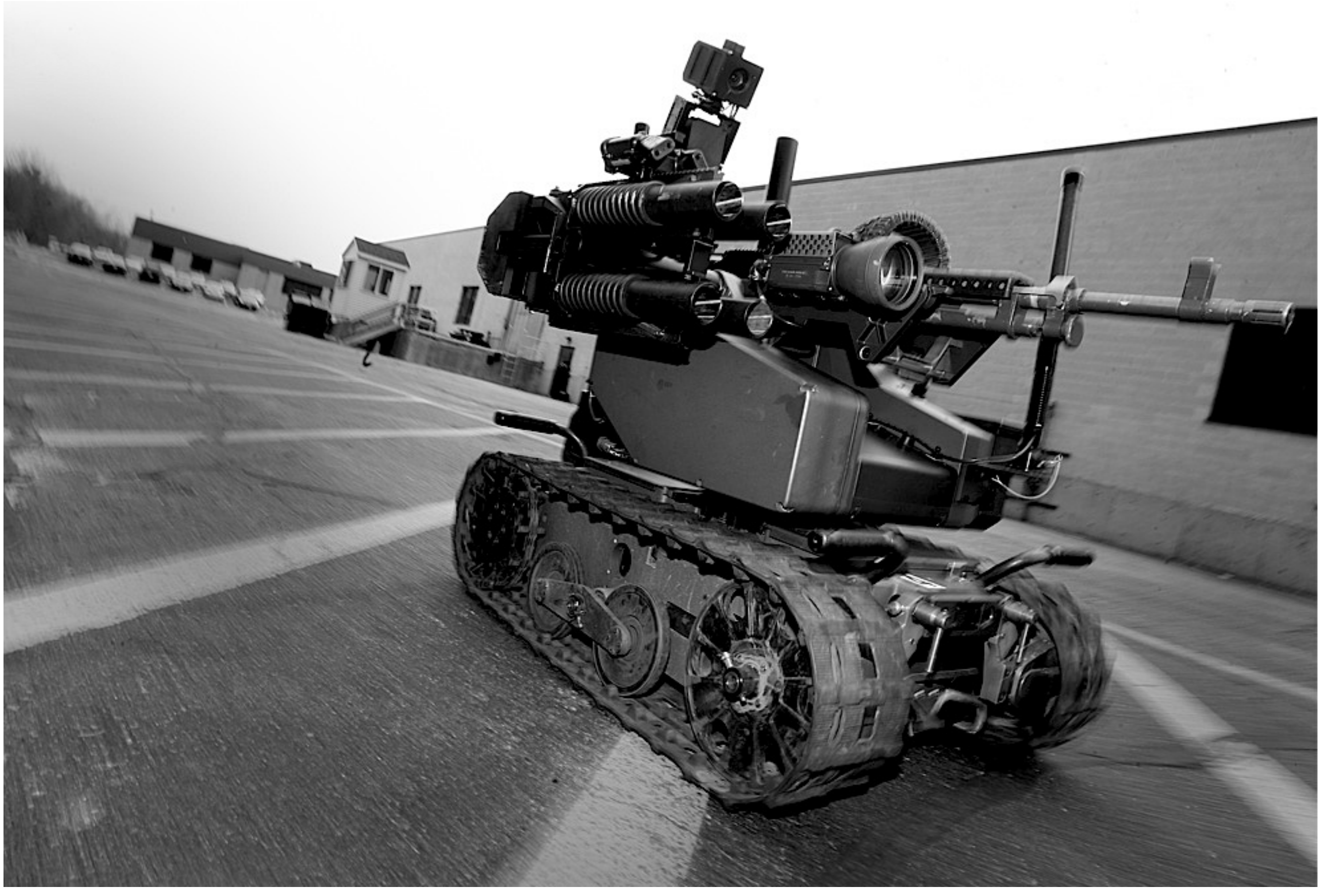

Fig. 1 The MAARS armed robot, showing multiple weapons mounted on the basic platform (photo from Boston Dynamics).

What would constitute a successful outcome of initial research in this area would be specifically a set of hard design principles/ requirements/guidelines/constraints, which would enable engineers to embed ethical principals of design and use into such a robot at an early stage in the Engineering Life Cycle. Note that although the scenario of use in for the robot is military, the same would apply for a civilian commercial domain

Furthermore the authors would postulate that the following types of outcomes need to be generated in parallel with the design guidelines referred to above;

- A generic, updatable State of the Art report including gaps and issues to be addressed
- A Context Base (legal/ethical/cultural): for example, for a given set of operational scenarios a set of significant legal, ethical and cultural requirements for the systems to be used in those scenarios based on operational need and the regulatory regime.
- Design guidelines and constraints on functionality that address these requirements.
- Exemplars for the application of the above
- A network of aware individuals that span a range of engineering, manufacturing, construction and other relevant domains (civilian and military) who may be able to contribute to the general research area and be willing to continue collaboration on future work on these issues
- A set of open proposals for future work, that will feed into future national / international programmes of research
- Evidence that industry and relevant government departments consider the findings to be valid.
- Seminars and presentations to Industrialists and interested parties in the UK, Europe, USA and elsewhere: examples in the UK would included the British UAV Association, the Society for Underwater Technology, the Unmanned Aerial Vehicles Association and the Motor Industries Research Association
- Publications in major, influential, journals concerned with autonomous systems

6. Acknowledgments

Thanks are due to the many people who have addressed these areas before us; we stand upon their shoulders to peer into the future.

7. References

1. McGeough, P., *America's deadly robots rewrite the rules*, in *Sydney Morning Herald*. 2010, SMH: Sydney, Australia.
2. Hallinan, C., *Death at a distance: the US Air War*, in *Foreign Policy in Focus*. 2007, Institute for Policy Studies: Washington, DC.
3. Clausewitz, C.P.G.v., *Vom Kriege (On War)*. 3rd ed. 1832, Princeton: Princeton University Press.
4. Hosken, G., M. Schmidt, and J. duPlessis, *9 killed in army horror.* I.O.L., 2007.
5. Sparrow, R., *Building a better WarBot: ethical issues in the design of unmanned systems for military applications.* Science & Engineering Ethics, 2009. **15**(169-187).
6. Sparrow, R., *Predators or plowshares? Arms control of robotic weapons.* IEEE Technology and Society Magazine, 2009. **28**(1): p. 25-29.
7. Sharkey, N.E., *The ethical frontiers of robotics.* Science :Vol. 322. no. 5909, pp. 1800 - 1801 DOI: 1, 2008. **322**(5909): p. 1800-1801.
8. Sharkey, N.E., *Death strikes from the sky: the calculus of proportionality.* IEEE Technology & Society magazine, 2009: p. 16-19.
9. Sharkey, N., *Grounds for discrimination: autonomous robot weapons.* RUSI Defence Systems, 2008(October): p. 86-89.
10. Arkin, R.C., *Ethical robots in warfare.* IEEE Technology and Society Magazine, 2009. **28**(1): p. 30-33.
11. Arkin, R.C., *Governing lethal behavior: embedding ethics in a hybrid deliberative/ reactive robot architecture*. 2007, Georgia Institute of technology.
12. Reggia, J., et al., *The Maryland large-scale integrated neurocognitive architecture*. 2008, University of Maryland: College Park, Maryland, USA. p. 50.
13. Moor, J.H., *The nature, importance, and difficulty of machine ethics.* IEEE Intelligent Systems, 2006. **21**(4): p. 18-21.