Autonomy's Troubling Questions

by Nick Hales

I was asked at a recent SCSC seminar whether I thought standards should be maintained in the face of the pressure by manufacturers to reduce standards so that they could compete, in the small aircraft business, with developing economies. My response at the time was, "Oh yes! It's the duty of every civilisation to raise others to their standards". This implies that we must persuade developing economies to adopt our rigid safety standards, but I realised later the statement is also fraught with the need for caveats. To use an orthogonal example, an environmental activist such as myself may say, "but I should add that in the case of primitive tribes such as the Andaman island dwellers or remote tribes in the Amazon, the standards we raise them to should be simple, such as providing free medical care where requested, leaving them to live as close to harmony with their wilderness homes as they wish". It is often true that we should make allowances and see the downsides of spreading a civilisation, or in the case of this essay, an ideology - autonomy - before we promote ourselves or impose.

This essay examines the ethics of autonomy as a developing technology. We must decide what is best of our safety-oriented behaviours to impose on the future of that technology without interfering in its purpose, i.e. to relieve mankind of the burden of labour. The particular focus of my discussion is lethal robots, so for the purpose of the argument the underlying principle is that it is morally permissible to resist aggression via force to achieve peace, the Falklands war being considered by the author an ethically reasonable example.

An Unbreakable Principle of Autonomy

In their joint work 'Formal Verification of Ethical Choices in an Autonomous System' [1], Dennis et al state what I will call the *Unbreakable Principle of Autonomy*: that "All participants in society are required to follow specific regulations and laws. An autonomous system cannot be an exception." Of course, many societies exist on earth and they follow diverse regulations. One can lose one's head for adultery in some countries but just one's partner in the UK. It is therefore essential to tie rules on autonomy to internationally-recognised standards such as the UN's *Universal Declaration of Human Rights*, thus protecting people from autonomous vehicles which may be used to try to deny those rights.

Among the Oxford Martin School's recommendations in the policy paper, 'Robo-Wars: The Regulation of Robotic Weapons' [2] is the following: "Ensure compliance with legal and ethical frameworks, guaranteeing that they are taken into account at every stage of the design process." That illustrates again that mainstream thinking considers current frameworks, such as the Declaration of Human Rights, as essential to create a code for autonomy, the framework, rules and regulations or laws for which are not yet clear.

One's initial thoughts on autonomy may be to consider some form of Formal Logic to control behaviour, as Formal Mathematics is used in the development of the design of safety critical software systems. However, one may also have to consider whether developments in Artificial Intelligence will enable future machines to act independently as moral entities. For some people, this brings to mind the frightening prospect of the *Terminator* films, where an interconnected system

of systems becomes 'self-aware' and then programs, via a system akin to the already emerging internet of things, all other connected machines to destroy mankind as the only threat it faces. While such may sound fanciful, the question arises, "As far as we can estimate it, what is the likelihood of this type of scenario?"

To illustrate how incremental increases in capability can create unexpected outcomes, it is only necessary to consider evolution. Meteorites have been found with amino acids preserved inside, even after falling through the earth's atmosphere and experiencing extremely high temperatures on their surface. Over a century ago, if one asked a group of people, "Do you believe that the fundamental building blocks of life can exist in rocks in space, and if so, are your notions so fanciful that you believe a rock could break up on impact and those building blocks (amino acids as we now know them), falling into a warm ocean or pool of water, could develop into a life-form as complex and capable of making ethical or non-ethical choices as yourself?", all or nearly all would have said "No". But that looks like a possible route for the beginning of our evolution. Therefore, one could conclude that once the reasoning of a computer connected to the Internet of Things becomes so fast and innovative that we cannot fully comprehend at any one instant of time how it arrived at its current state, or whether it is developing its own sub-routines to perhaps defend itself from viruses, worms and Trojan horses etc., then there must be a perhaps significant probability that such a similarly inanimate building block could develop consciousness. We cannot escape this by believing that humans are differentiated from the rest of nature – or indeed of matter – by having 'souls' or something equal, since observation of animal behaviour since the 1960s leads us to the conclusion that, in some cases, they too have some form of ethics, reasons for conflict over resources, resolution of differences capacity, friendship, identification of enemies etc. The possibility that apparently inanimate rocks from outer space contain amino acids which are responsible for life on earth, illustrates how things can develop from next to nothing, biologically speaking.

Ethical Decisions By Machines

Continuing with the point made in the paragraph above, self-protection by an autonomous vehicle could be permitted until a decision is to be made which will affect a human life, or upon which a human life may depend. At that point, the decision the machine makes must be identical to that which a consensus of humanity would make in that position but excluding the legal principle of 'capacity' since the machine cannot be claimed to have capacity. The principle of capacity enables, for instance, a decision to be made about whether the life of a mother is more important than the life of a child she bears, by adults with capacity but not the unborn child. For instance, abortion is often justified as being within the woman's capacity, so she takes (moral) precedence, but in Roman Law, as a last resort, the child could be removed from a dead or dying mother: the woman of course would then die but the child live - the child took precedence. (The birth of Caesar this way is almost certainly a myth since his mother is reported to have lived long after his birth, an impossibility then). However, in the argument here, the survival of the machine is irrelevant, since it is non-human and therefore has no capacity. Only the non-combatants can be counted as having capacity which the machine should consider. Since the mission obviously has a high priority and may involve the death of opposition combatants, the machine would not be required to consider the capacity of those combatants, since their involvement in a

combat scenario effectively removes from them the capacity they would have in normal everyday life. The precedence rule therefore applies only to those who would be affected by a bad decision and are not opposition combatants. It effectively preserves the sanctity of life and possibly the environment if one takes the principle of 'Ecocide' (Higgins [3]) on board. The machine may sacrifice itself or terminate its mission to preserve the life of its masters, allies and non-combatants.

The question could then become, "Is there an intention to consider non-combatants as collateral damage?" If there is, then the machine, it is suggested, would have to be fully aware of the complete environment to test whether or not a 'shot' could be taken a few minutes later when there are fewer non-combatants in danger. It would also need to be able to calculate the numbers of non-combatants likely to be affected as collateral damage, since it will clearly be more likely to be acceptable to the public if four non-combatants got caught up in an explosion than if a crowded market place is hit and 200 innocents die in the explosion and any subsequent conflagration which rages through the market. This will be so difficult that it is considered unlikely ever to be permitted. The difficulty is that the decision as to whether to attack when the death of non-combatants is inevitable is, in current circumstances, a command decision passed to a combatant, (a classic example being the decision of Commander Dick of the London Metropolitan Police to permit termination of what proved to be the life of the innocent Jean Charles de Menezes on the London Underground in 2005). Thus, if a machine was given the power to make that decision then the machine – rather than the front-line soldier - would effectively become the Commander, and by the primary rule above a machine must not be allowed to become an exception to the principle that "all participants in society are required to follow specific regulations and laws". By becoming a front-line soldier with high command control it would be indeed doing that.

The need would be for the unmanned system to have recognition systems which allow it to perform various judgemental functions without human intervention, i.e. to know the difference between combatants and non-combatants, to calculate the numbers involved and to be able to predict whether or not a more opportune moment is likely to arise. At the beginning of the James Bond movie, *Skyfall*, M orders an agent to take a shot, knowing James Bond could be hit, simply because it is the best chance, as she sees it, to kill the enemy agent Bond is in close combat with. It is argued here that society cannot ask a robot to play the role of M in similar scenarios. In any case, it could be argued that M made the wrong decision, in which case programmers could get blamed for acts which commanders would have been responsible for, freeing commanders from guilt but making system designers culpable. 'Friendly Fire' casualties due to poor robot engineering will be unacceptable to the public.

Ethics, Cyber and Stealth

An ethical autonomous system would need to be able to discern what choices it has and make the best decision. If no option is available which fulfils its programmed 'ethical' criteria - if indeed a program can be designed to answer the questions posed by this complex and context specific problem - it is thought by some that it may then examine options in order of least "unethicalness" and select the best. But how that could be achieved is open to question. How can safety be maintained if security is not so powerful that guarantees that the system cannot be penetrated can be provably factual? Option

programmes may be the first point of attack by cyber agents to turn the machine against its masters. Also, what would happen if given the above, the least unethical options available is something completely unacceptable to humanity, despite other options being worse? The machine must clearly also be capable of knowing what the public would accept, given the risk that its priority target stack could be altered by an enemy.

A large international company involved in the development of an advanced robotic weapons has already made it clear that, with regard to unmanned aircraft systems, they believe and design to the principle that there will always be a need for a human in the loop, in particular regarding any use of weapons, both now and in the future. It remains to be clarified at what stage of the loop exactly the operator is involved. This was also the opinion of an international group of safety-aware engineers meeting in 2015 near Washington which the author attended. However, in case politicians change their minds, it is certain that many companies' future-proofing of unmanned systems technology will involve the inclusion of options for complete autonomy. In flight, for example for evading missile attacks, this is perfectly acceptable, since it would be impossible for a remote human operator to manoeuvre the vehicle in a sufficiently agile manner to achieve such objective-related mission critical activities. However, in a dog fight, could there be certainty the UAS would not attack its own manned platforms in the confusion? It would therefore appear that stealth to evade such combat problems would be a preferred option to autonomous combat. In-between options exist, where a human operator may be actively involved in particular steps of the targeting process while other steps are left to the robot, once it has been programmed accordingly. But if control and communication are lost, the intended intervention may not take place.

One important factor is that some extreme politicians may, simply to gain popular support, insist on ethical controls being lifted. This would imply over-ruling all engineering and corporate advice and demanding fully autonomous systems, saying damn to the risks and the potential for unwelcome tit-for-tat responses by less scrupulous regimes, instead of pursuing international agreement through the auspices of organisations like the UN.

Three Principles

Robo-Wars [2] cites three principles for developers, users and regulators: "Recommendations for States","Recommendations for manufacturers and the military" and "Recommendations for States and the military".

The principle for regulators or government has been dealt with partly in the discussion on ethical frameworks above. For the users or military/ police forces, the essence of the recommendation



is that requirements should include allocated responsibility and the enabling of over-ride by operators. However, as discussed above, a chicken-and-egg situation arises here, since the requirement must be bound by regulators no matter what the military wants. Beyond these Oxford Martin recommendations, there is also a case for an automatic "return to base" program to be triggered if communication is lost (this principle might also be useful for hijacked planes, by the way, where the pilot immediately on realising a suicide, bomb or hijack attack is underway, presses one single button which initiates a programme that takes the plane to the nearest safe airport automatically and which cannot be over-ridden).

The last of the three, "Recommendations for states and the military" for reasons discussed, must be given priority because otherwise manufacturers may produce whatever anyone requests, even if this lies outside the envelope of developing ethics. It can be summarised thus: "At an international level, develop standards for the deployment of robotic weapons, specifying and delimiting the contexts and purposes for which robotic weapons can be used, preventing use outside legal paradigms provided by international, humanitarian and human rights law", the former of which requires distinction and proportionality in war, the latter of which concerns law enforcement or police activities.

If the three principles can be summed up in one sentence it is this: "Autonomy is used to enhance human decision-making, not substitute for it".

The Lack of Chivalry

Most people will have thought at some point about uprisings, as weapons become more sophisticated, "They may be the losers and I would not want them to win, but one cannot deny their bravery". The problem of autonomous systems is that, as they increasingly become able to operate without involving the personnel of the military force that uses them, the former codes of chivalry (reliant on ideas of loyalty, fair-play and honour) no longer have any meaning. Armed forces may be seen as disabled by their use of robots in the eyes of those who oppose them in asymmetric conflicts. These insurgents and terrorists can capitalise on this model, attracting sympathy where otherwise their cause would be despised. Military doctrine in the west already holds high principles, so though it has been suggested in [2] that "responsibility and integrity" could be virtues ascribed to non-combatant Remotely Piloted Air System operators, where chivalry and bravery cannot. This is unlikely to have much traction, and rotation through the role by former front-line combatants who have already earned their stripes and medals is more likely.

Conclusion

The future can look bleak or rosy, depending on one's state of mind and the information one has received. A rosy future may imply a shared sense of global "civilisation", but would you want it policed by robotic systems, with all the questions of ethics that then arise and the potential for unsafe operation or penetration of the system by cyber-attacks? A bleak future may entail endless conflict against asymmetric forces or even ludicrous virtual total destruction initiated by advanced civilisations, armed with unthinkable power, engaging in relatively trivial disagreements or, worse still, due to cyber-attacks by terrorists on nations otherwise at peace. In any of these events, forethought by engineers in producing safe robotic systems operating under moral paradigms of behaviour, could be the difference between the possibility of peace or total destruction.

I will close by offering one of the most moving and concerning statements given by a scientist recently. Professor Brian Cox, talking about life elsewhere in the Universe, made a comment to the effect that, because we have developed from cavemen to advanced armed civilisations so relatively quickly in astronomical terms, our intellectual power has not caught up with our power to destroy. We are like children operating machine guns. If this has always been the case, wherever intelligence evolves in our universe, it may be that every advanced life form vanishes in an orgy of self-destructive violence soon after evolving and that therefore, despite the billions of possibilities, we are alone in the universe. Earth-based life forms may be doomed never to communicate with intelligent life on other planets because we cannot make the necessary ethical decisions to enable long term survival. This is very sad. If we are to move safely forward to the future we must maintain our humanitarian standards and encourage others to adopt them. Developing new codes of practice for autonomy is therefore essential: future generations of technology, and of humans, depend on them.

References

[1] L. Dennis, M. Fisher, M. Slavkovik and M. Webster 'Formal Verification of Ethical Choices in Autonomous Systems' , Robotics and Autonomous Systems vol 77 (2015), pp. 1-14 http://www.sciencedirect.com/science/article/pii/S0921889015003000

[2] A. Leveringhaus and G. Giacca, *Robo-Wars: The Regulation of Robotic Weapons*. Oxford Martin School, University of Oxford, 2014. http://www.oxfordmartin.ox.ac.uk/news/201412_Robo-Wars

[3] P Higgins http://eradicatingecocide.com/

Until his recent retirement, ***Nick Hales*** *was a Senior Engineer at DE&S, specialising in critical avionics systems.*

PLEASE NOTE: The views expressed in this article represent the fruits of Nick's own thinking, and should not be ascribed in any way to his previous employer, or to UK MOD.

We are pleased to announce a new service for SCSC members and member organisations: SCSC Announcements. From March 2019, we will be happy to take suitably-worded announcements of interest to members (for example, of forthcoming events, publications, calls for tender or vacancies) and advertise them on the SCSC website. The link will be provided in *Safety Systems* and will also be clear from the home page. For further details, please contact Alex King: *alex.king@scsc.uk*

