Could machines be responsible for a major disaster?

Danny Weston
Centre for Automation and Robotics Research, Sheffield Hallam University
(D.L.Weston@shu.ac.uk)

The government of South Korea plans to have a service robot in every home by 2020. Robots already available for such roles can do many tasks, from cleaning the floor to feeding someone breakfast. Meanwhile, on the border with the North, sentry robots are deployed with complete autonomy to kill approaching human-sized targets. Is this, to paraphrase a somewhat hackneyed expression, merely the difference between a sword and a ploughshare? No. We customarily think of robots as such tools, yet the engineering decisions behind both the cleaning robot and the sentry are inextricably linked. Values are embedded in both, and they are also both capable of actions independently - whether semi or completely autonomously - in a way that a sword or a ploughshare is not.

Some distinctions

As Wallach and Allen put it[1], (ro)bots extend "the field of computer ethics beyond concern for what people do with their computers to questions about what the machines do by themselves". Machine ethics of course, goes beyond the raw actions a device is capable of; it also includes all of the assumptions and expectations that are built into its very design. At this level of understanding, one can reasonably make the case that any machine has 'values' built in that predetermine its physical characteristics, capabilities and range of actions and uses. Many of the latter may not be anticipated - creating a further conundrum (see the case of the Roomba below).

Intuitively, this may not necessarily be clear or obvious. However, it is straightforward to illustrate in the case where ethical design concerns are *not* explicitly considered in designing and building a machine: If a designer chooses, consciously, not to consider the consequences of decisions made or actions carried out, by their machine, this is *still* embedding a certain set of values into the system.

The major pioneer in considering what sort of 'values' our technologies may have embedded in this way was probably Lewis Mumford, an American historian and philosopher working in the early to mid 20th century. He referred to such embedded values as 'technics'[2], making a primary distinction between 'authoritarian' and 'democratic' 'technics'. These focused on not just the technologies themselves, but also the practices surrounding them and applied to technologies more generally, not just complex machines. A more modern iteration of Mumford that makes more sense

[1] Wallach, W. and Allan, C. (2009) 'Moral Machines',OUP, p.6
[2] Mumford was discussing 'technics' as early as the 1930s, however for a good summary of his views it is worth consulting 'Authoritarian and Democratic Technics', *Technology and Culture*, Vol. 5, No.1 (Winter, 1964), 1-8.

in the 'information age' is the idea of 'enclosing' and 'disclosing' dynamics[3]. Appropriately classified technologies would help to withhold and contain information or produce and transmit information, respectively. These notions are obviously useful for information and virtual based technologies, but less so for actual machines.

To dig a little deeper into more specifically useful terms, a sharper distinction is made by using two scales - from low to high autonomy on the one hand and from low to high 'ethical sensitivity' on the other. Two examples from Wallach and Allan[4] can be borrowed to illustrate: i) An autopilot on a plane can be regarded as a system with substantial autonomy, but little ethical sensitivity. With this system, an aircraft can fly automatically in a range of conditions with little to no human supervision. While some level of 'ethical sensitivity' is built in - for example, banking less steeply than it could (or would be optimum) to ensure passenger safety, it has no sensitivity to unusual conditions - for example a sick passenger on board that may require even more gentle manoeuvring. ii) A computerised decision support system (an "expert system") such as 'MedEthEx'[5] has low autonomy but a greater degree of 'ethical sensitivity' (at least in a very circumscribed area of ethics). While it would be rated higher than a passenger jet autopilot, such an example does not score highly on the 'ethical sensitivity' scale because it tends to be used for educational more than operational purposes. It could barely be described as an 'advisory' system, as it deals with general principles, rather than analysing specific cases and responsibility still lies with the practitioner.

If we assume that autonomy itself is going to increase in machines, ineluctably, then our focus should then be on how to improve their 'ethical sensitivity'. Some recent attempts have been made to make such advancements in the realm of military robots, working towards creating an "ethical governor" to ensure they follow the Rules of Engagement and Laws of War[6].

So we have a two dimensional scale on which to quickly assess a machine or system. Can we get any closer to a more practical taxonomy though? James Moor[7] proposes categories of ethical agents. From lowest level to highest - 'ethical impact agents', 'implicit ethical agents', 'explicit ethical agents' and finally 'full ethical agents'. Moor argues that the practical end-goal of machine ethics should be 'explicit ethical agents'. The first - 'ethical impact agent' is simply *any* machine that could have an ethical impact - which is virtually any machine one could think of, and most definitely any that are able to move or have accessible moving parts (for example, cars). The next level 'implicit ethical agents', concerns machines where effort has been made to

[3] See for example May, C. (2002), 'The Information Society', Blackwell.
[4] Wallach and Allen (2009) (pp.26-7).
[5] See 'MedEthEx Online' at http://webcampus.drexelmed.edu/medethex/index.html
[6] See for example, the work of Ronald Arkin, 'Governing Lethal Behavior in Autonomous Robots', (2009), Chapman-Hall
[7] See for example Moor, J. H. "The nature, importance, and difficulty of Machine Ethics", *IEEE Intelligent Systems Special Issue on Machine Ethics*, vol. 21 ,no. 4,pp. 18–21,2006.

avoid negative ethical effects through addressing critical safety and reliability issues in the design (elevators for example have many redundant safety systems). The third category - 'explicit ethical agents' are machines that reason in some way regarding ethics in their programming. There are multiple approaches currently being researched in this area ranging from installing deontic logics through to using learned (evolved) behaviours. Finally is 'full ethical agents', which refers to agents that act on a more-or-less human level of intelligence and comprehension; something currently far beyond our technical means, even in the most complex simulations, but a distinction worth making. Some computer scientists and philosophers doubt even the possibility that this could be achieved in the far future. However, the question isn't 'can we make truly ethical machines?' - One of the goals of those focused on the idea of an A.I. or technological 'singularity'[8]. Rather, it is, 'can we make machines, whose behaviours will have ethical consequences, behave in an appropriate manner?' Moor's argument is that in making these distinctions we can clearly make the case that our aim should be towards 'explicit ethical agents' – we plan and design ‘as if’ our systems will be able to make at least *some* ethically significant decisions.

The Roomba example.

The Roomba is a good example of a popular service robot. It is used by thousands of people to automatically clean their floors, usually whilst they are out of the house. One of the first problems identified in its initial development was the possibility that it may reach the edge of a stairwell and fall down. While there was a concern that the robot may inadvertently damage itself, the main issue was that it may fall and strike a human or pet. A piece of solid metal, ceramic and plastic weighing in at 6.5lbs and tumbling down a staircase could cause a serious injury.

This was solved by the careful use of sensors, allowing the robot to detect when it had reached the edge of a surface with a sharp drop. However, reports later surfaced that the robot had other, unexpected behaviours that in their own way, however seemingly innocuous at first, could actually be quite dangerous. Several users have reported that they had found their robot in unexpected places, and in some cases it had closed a door. Such a relatively simple and ostensibly benign machine as the Roomba can still make a sufficiently hefty doorstop[9]. It is obvious that this could be a serious risk, especially for vulnerable people, or in the case of a fire.

Thus the Roomba, whilst satisfying the criteria for Moor's classification of 'implicit ethical agent' is nevertheless able to behave in ways that have a significant ethical impact regardless of the fact that the machine has no ethical decision making

[8] See for example, the Singularity Institute's discussion of 'Friendly AI' - http://www.singinst.org/ourresearch/publications/what-is-friendly-ai.html
[9] See for example, discussion on this on Slashdot http://tech.slashdot.org/article.pl?sid=05/02/01/1536256 - last accessed Nov 9th

capability (and as far as its designers were concerned, needed none in the first place).

With the law of unintended consequences already at work with a lone – and relatively simple – device such as the Roomba, it behoves us to think ahead to the near future. The move to ambient intelligence and ubiquitous computing in the home[10] could easily multiply unexpected behaviours at an alarming rate. Consider scenarios where the Roomba is one of many devices interacting? Multiple devices linked together by wi-fi could result in some novel emergent behaviour, especially if any of them are able to dynamically download soft/firmware updates and influence other devices. There are already enough conflicts, and bizarre behaviours, caused by two or more pieces of software from different vendors conflicting on the same PC.

There is of course always the possibility of outside interference from humans. The story on Roomba behaviour cited above is followed by numerous comments from people claiming to have used relatively simple sensing technology to terrorise friends and colleagues with mechanically based practical jokes (such as a hidden device that produced chirping noises 20 minutes after the lights had been turned off). Numerous household devices linked together with wi-fi or bluetooth networks will seem like very tempting targets to curious hackers, especially if one is able to cascade behaviours from one device to another (changes to a thermometer in a house for example could lead to all sorts of changes in other devices in a 'smart home').

Whilst these scenarios may seem relatively limited, or even humorous, it is a straightforward process to expand such thinking to interlinking machines and systems with massive impacts on a national, or even international scale. Wallach and Allen open their book *Moral Machines*, with just such a scenario[11], an adapted and abbreviated version of which is presented below:

Abbreviated version of the Wallach-Allen scenario:

The future scenario starts in the U.S. on a day when electricity demand is expected to be high. Rising energy costs have made speculators drive up the prices of futures, and also increased the use of automated trading systems to profit from miniscule short lived variations in prices.
At 10.15 a.m. the price of oil drops slightly in response to the news of the discovery of large new recoverable reserves. Software at an investment bank calculates it can turn a profit by emailing a quarter of its customers with a buy recommendation for oil futures.
Unfortunately the instruction to buy is taken up far more enthusiastically than expected. Investors used to seeing the price of oil continually climb react to this news, pushing the spot price of oil beyond $300 / barrel.

[10] Gates, Bill, 'A Robot in Every Home' in *Scientific American*, Jan 2007, pp.58-65
[11] Adapted from the more detailed scenario in Wallach and Allen (2009), pp4-6

It is now 11.30 am on the East Coast. Temperatures are rising more rapidly than predicted. Software controlling the power grid determines that it can meet the extra demand while suppressing the cost by using the coal-fired, rather than oil-fired generators.
One of the coal-fired stations suffers an explosion while running at peak capacity. Cascading blackouts affect the power supply for half of the East Coast. The New York Stock Exchange is affected, but not before regulators notice that part of the rise in oil prices was due to a shell game between automatically traded accounts at one large investment bank. As the news spreads, it is clear that prices will fall dramatically as soon as the market reopens. Meanwhile the blackouts continue to spread.
Detecting the expanding blackouts as possible terrorist action, security software at a major airport sets itself to high alert and applies biometric matching criteria that will increase the number of people flagged as suspicious. The software has no built in mechanism for weighing the benefits of preventing terrorist action against the inconvenience that may be caused for thousands of people. With the enhanced criteria in effect, a cluster of five people due to board a flight to London are identified as suspicious. This large concentration of suspects on a single flight causes the program to trigger a lock down of the airport.
As a result, sentry guns installed on the U.S.-Mexican border receive a signal from Homeland Security that places them on red alert. In such situations they are programmed to act autonomously, without human oversight. One of the sentries targets a jeep approaching from off-road, destroying the vehicle and killing three U.S. citizens who were on a 'technology free' camping trip and didn't receive the alerts sent out via the internet and SMS that the Homeland Security threat level was red and strict protocols were in force on the borders.

The above example may strike some as somewhat fantastic. Yet real examples of parts of the hypothetical scenario already exist. For example, in 2003 there was a cascading failure of the power grid in Ohio, in the U.S. - the "worst outage in North American history"[12]. Subsequent investigations found that one of the main culprits was a subtle race condition bug in software used across the grid. Similarly, data mining and profiling software has become notorious for producing a large amount of false positives, particularly in the arena of security and anti-terrorism[13].

On the more prosaic level, a staple of debates using ethical dilemmas have been the 'Trolley cases', introduced by philosopher Phillipa Foot in 1967[14]. These are thought experiments concerning hypothetical automated tram systems that have to make - what appear to be on the surface - simple decisions about their route, speed and so

[12] See for example, the reporting on this in The Register - 'Tracking the Blackout Bug' http://www.theregister.co.uk/2004/04/08/blackout_bug_report/ - last accessed Nov 27, 2009.
[13] Schneier, Bruce, 'Why Data Mining Won't Stop Terror', *Wired*, March 2006.
[14] Foot, Phillipa, “The Problem of Abortion and the Doctrine of the Double Effect.” - reprinted in numerous places, an online version (with commentary) is available here: http://www.econ.iastate.edu/classes/econ362/Hallam/Readings/FootDoubleEffect.pdf

on. The examples over the years while having increased in sophistication and subtlety are usually some variation on how the computer should weigh the value of people's safety, or even lives, given conflicting options (e.g. someone on the track versus the passengers it is carrying).These examples are highly relevant now given the increasing number of automated tram systems now in operation across the world (a relatively recent crash on the Washington Metro system was at least partially blamed on automated systems[15]).

The idea of 'responsibility' feeds in here in an important way – for it involves going beyond a simple technical fault, or a component failure (though it is inclusive of those). It also includes situations where decisions are *explicitly* made in the programming. A device could have very simple goals, but if it has adaptive capabilities, the consequences could be near impossible to predict. A paper by Stephen Omohundro[16], for example, points out how an artificially intelligent machine, whose primary purpose is to play chess could resist being switched off, and behave in many unexpected ways.

Whilst at a philosophical level the issue of 'responsibility' is extremely thorny, it is easy to bring it back down to a practical level very quickly: Commentary[17] on the Washington Metro crash on how to determine responsibility is very apropos in this regard: "The modern world is full of movements that are overly concerned with motivations, and it is passé to worry about whether whatever cause you're espousing will actually accomplish the grand goals that are claimed for it. Bluntly put, people are too concerned with other peoples' wishes, which are none of their business, and not enough concerned with other peoples' competence, which is very much a legitimate concern. You'll find that for things that really matter like not having train wrecks — people pretty much all want the right thing already."

Conclusion

Moral agency is often equated with moral responsibility. This is something that many researchers in the field(s) of machine ethics have been at pains to separate clearly. And it is - at least in most cases - applied in the sharp end of practical situations. Predator drones used by the U.S. military for example, always require a human (usually an entire chain of command) to authorise kill shots. There is no guarantee this will always remain the case however and some systems already have autonomous killing capabilities, such as the Sentry mentioned at the beginning.

[15] 'Washington Metro Delayed Upgrades', *The Wall Street Journal*, June 24, 2009 http://online.wsj.com/article/SB124573949695640729.html last accessed Nov 28, 2009

[16] Omohundro, S. M. 2008a. The Basic AI Drives. In *Proceedings of the First Conference on Artificial General Intelligence*, 483-492. Amsterdam: IOS Press.

[17] From 'The Foresight Institute' - http://www.foresight.org/nanodot/?p=3123

Ideally, moral philosophers specify that this separation is maintained, following something like Moor's classification of 'explicit ethical agent'. Even were this ideal realised however with every single component, machine and program that could engage in, or influence, ethically significant behaviour, the examples ranging from the humble Roomba to the Wallach-Allen disaster scenario demonstrate that unexpected behaviours and outcomes are possible, if unlikely. We are therefore returned to the conundrum of balancing the precautionary principle against making social, technological and scientific progress.

We are used to making just this separation with regards to humans with particular characteristics, for example children, or the mentally incapacitated, who it is argued, can carry out morally significant actions but to whom we cannot ascribe true moral agency as they are not fully conscious of what they are doing.

Such a balance is only possible on the basis of both engineers and philosophers acknowledging these issues in the first place. Progress is at least possible given this awareness. Without it we could quickly stumble into hostile territory. There is a danger, for example, that using tools such as MedEthEx could quickly become like crutches, where their users forget that it is more an educational and informational device than an advisory system. It isn't designed to share moral responsibility with the medical practitioner, yet some may be willing to trust it more than the 'fallible' human. So if we remember anything, it should be the purpose for which anything is built; complex machines are expendable, if massively value laden, tools. Humans are not.