Rise of the robots: Rethinking ethics, trust and responsibility in the age of autonomous machines

Darren Dalcher

Hosting the Olympic Games is viewed as an opportunity to celebrate achievements and showcase new technologies. In October 1964, prior to the Tokyo Olympic Games, The revolutionary Maglev bullet train was unveiled as the fastest train in the world. Throughout the next 51 years of operation, the trains have transported over 10 billion passengers, often travelling through typhoons and earthquakes using sophisticated detection and alarm systems without suffering a single incident involving loss of life.

To celebrate the return of the Olympic Games to Tokyo in 2020, visitors will be introduced to a robotic experience. Over one million visitors to the Odaiba district will be hosted in a futuristic village, where robots will hail taxis, fetch luggage, administer check-in desks, operate hotels, offer instantaneous translation services and ferry visitors to their destinations. According to the Japanese Prime Minister, Shinzo Abe, Japan is even planning to stage a Robot Olympics alongside the summer games. Whilst the games may once again offer a glimpse into the future of a new technology, it is a controversial future that demands a greater trust in autonomous robots and their ability to make safe and ethical decisions.

1. Why the driver cannot be blamed

Modern technological advances continue at an unprecedented pace, proudly displaying greater autonomy and decision making skills embedded into an expanding range of technologies including artificially intelligent robots, self-driving cars, delivery by drones, ubiquitous mobile supercomputing, implantable technologies and smart cities. As the new technologies are deployed to undertake tasks as diverse as educate, entertain, eliminate enemies, deliver parcels, drive, guide, satisfy and inform us, they also take over a growing number of repetitive and dangerous duties and chores that were previously handled by humans. However, when such technology is given full autonomy for making decisions it can also play a part in introducing a new kind of computer-assisted error, where a system designed to make us safer is directly responsible for causing an accident. In abrogating responsibility for mundane decisions to new technologies, we are increasingly relying on their ability to deal with risk, uncertainty, ambiguity and the greater unknown.

Japan has long viewed robots as a major pillar of its economic growth strategy and an important aide for a rapidly ageing society. In preparation for the 2020 Tokyo Olympics, *Robot Taxi* is field-testing its new driverless taxi service. Starting in March 2016, 50 residents of Fujisawa, a large coastal city south of Tokyo, known as Japan's first sustainable smart town, are regularly being driven between their homes and the city's supermarkets, some two miles away. The autonomous cars combine GPS, radar, stereovision cameras and image analysis systems to navigate around town. Successful trials are expected to lead the way to the use of the thousands of robot taxis to ferry spectators around the 2020 games venues.

Capable robots with 360-degree vision, full awareness of the environment and perfect driving skills may yet force a redefinition of humans in cars as cargo. But until such precision instruments replace all drivers, driving will require interaction with other road users, which may often mean flashing lights, gesticulating or making eye contact. However, as future human cargo, we should all be interested in the choices made by autonomous driverless cars. As the car you are transported in careers towards a junction at the bottom of a hill, whilst gathering speed as the brakes fail, would you expect your autonomous taxi to come to a halt after running down the old lady at the bus stop, or would you prefer that it headed straight for the five young men in the open area by the town square? Would the decision change if the person in the bus stop is your boss, or one of the young men is your youngest son? At any rate, should the passenger care as long as they are kept safe?

2. Whose moral reasoning?

Moral philosophers have utilised thought experiments to debate choices and uncover the circumstances under which it is acceptable to harm others. Dilemmas, involving technology and a choice between different options question whether it is acceptable to divert a trolley bus about to kill five people towards a single individual in order to save the five, and whether pushing a single 'fat' man onto the track to save the five is equally acceptable. Different formulations can thus be used to highlight the modern quandaries created by robotics and new technology advances.

The human responses to such trolley dilemmas typically highlight the role of consequentialist vs. categorical moral reasoning. Consequentialists focus on the results of action to determine if it is right or wrong, implying a need to maximise good, or 'less bad', results. Categorical moralists take issue with each act, investigating its appropriateness, arguing that some actions are categorically wrong. The former position locates morality in future consequences, while the latter takes issue with the act itself, thereby locating morality in certain duties and rights and dealing instead with rules and absolutes.

Emphasising harm avoidance, harm minimisation, or utilitarian maximisation may lead to very different outcomes. In programmed artefacts, the system of preference needs to be coded and acknowledged, invoking a particular method of reasoning about safety, responses and consequences. So in June 2010, when the US Military lost control of a helicopter drone for over thirty minutes and twenty-three miles as it swerved towards Washington DC, potentially threatening the White House and other civil and military assets in direct contravention of established airspace restrictions, relevant agencies would have benefitted from knowing whether it was armed with missiles or parcels, as well as what moral system it might be deploying.

3. From Frankenstein to Asimov

Consideration of the role of ethics in new technologies is not new. Fear of dealing with robotic creations and their unpredictable behaviours has repeatedly featured in literature. Mary Shelly's Frankenstein, often used as an allegory for the folly of scientific experimentation, actually tells the story of an 'assembled' powerful creature, capable of extreme and destructive violence, who also learns to speak, secretly cares for a poor family, reads literature and yearns for a soul mate, and his struggle to reconcile power, autonomy and feelings. It can also be read as a commentary on an irresponsible creator who fails to recognise and embrace his responsibilities to his creation and to society at large.

Issac Asimov formulated the Three Laws of Robotics:

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2. A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.
3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws.

Asimov's writing is concerned with the safe behaviour of autonomous robotic machines and their greater impact on individuals and society. The fictional stories explore the dilemmas of unexpected events, counter-intuitive behaviour, unexplored boundary conditions and the unintended consequences of applying such laws.

Ethics is often slow to catch up with technological developments. A key question that emerges from the writing of both Shelley and Asimov is whether machines can act as moral agents. If robots are to take on added autonomous roles, they must be programmed with moral decision-making responsibility–but whose morality do they take on?

Shelly's dystopic tale of demonised technology emphasises the result of execution gone awry, in a consequentialist tradition. Meanwhile Asimov's use of the three laws as a literary device exemplifies the difficulty in enforcing a categorical value system through the use of absolute rules and prohibitions.

4. Humans, Feedback and Staying in Control

In many human endeavours intelligent automation is replacing some of the tasks and roles traditionally performed by human agents. The key reasons for employing intelligent technology is superior computational capability coupled with elimination of human error, and the reduction in work overload and lack of dependability. However replacing humans may increase system vulnerability, especially in reference to unanticipated perturbations, which cannot be foretold or specified.

Reliance on autonomous machines requires total trust in the ability of the system to make safe or rational decisions. Replacing human decision makers with a 'responsible' mechanical alternative would therefore pre-suppose that all possible failure modes had been foreseen and specific actions were included to mitigate their effects. Given that many contexts cannot be fully specified in advance, solutions often need to evolve through experience by dynamically consulting domain-specific experience. Proponents of artificial intelligence, a field which studies how to create machines capable of intelligent behaviour, contend that autonomous cars are able to learn from incidents and the resulting corrections and changes then apply to the entire class, not just to a single autonomous agent. In other words, all cars become smarter following an accident. Yet, overreliance on technology often results in ignoring the human element.

It is expected that visitors to the 2020 Olympic Games will be staying in robot-staffed hotels. The first such hotel, Henn-na, said to be the first hotel totally run by robots, opened in Nagasaki in July 2015. The hotel is operated by robot receptionists (with a choice between English-speaking dinosaurs or Japanese-speaking female androids), robot porters, and other electronic creatures, coupled with facial recognition technology and a multitude of sensor panels. Check in is available from 3pm; however, visitors who arrive early and attempt to engage the robots will encounter a human who comes out of his small room to announce that the machines will become operational at 3pm. Even a complete system, may thus require occasional intervention.

Ultimately, if a function is to be automated then the system must be supplied with enough variety and control to cope with any situations that might arise. In order to generate suitable control responses to address unexpected conditions, every controller must be provided with:

1. Sufficient control responses, and;
2. Decision rules for generating all the control responses, or;
3. Authority to become a self-organising system in order to respond to unexpected events, or;
4. A resident human ready to apologise for lapses.

5. Should designers anticipate surprises?

Inherent uncertainties and unexpected external conditions can lead to surprises, often necessitating an urgent intervention. Yet, adding interventions may itself lead to complications:

On June 1st, 2009 an Airbus A330 equipped with the latest 'glass cockpit' controls entered an aerodynamic stall from which it could not recover and crashed into the Atlantic Ocean, killing 228 passengers and crew members. The wreckage of the Air France flight from Rio de Janeiro to Paris was discovered five days later near Saint Peter and Saint Paul Archipelago in the central equatorial Atlantic Ocean. The accident report concluded that the crash occurred after temporary inconsistencies between the airspeed measurements caused the automatic pilot to disconnect. Crewmembers struggling to regain control in a sudden emergency misread the situation and reacted incorrectly, ultimately causing the aircraft to enter the fatal stall.

Imposition of new technology can also change the balance in the environment itself and introduce potential breakdowns in communication. Indeed, Professor David Woods, an expert on human interaction with technology, suggested that automation actually makes flying more difficult for pilots. British aviation expert David Beaty also documented typical automation errors that precipitated accidents while other aviation experts [1]observed that the modern cockpit solved a great many problems, but created some new ones as well, a direct parallel to driverless cars. David Beaty believes that pilots are increasingly being pushed out of the control loop. In the Airbus accident, bringing pilots back into the system in an emergency may actually have escalated the failure scenario.

Events pertaining to hazards interact with psychological, social, institutional, and cultural processes in ways which heighten or attenuate perceptions of risk and shape behaviour. Professor Charles Perrow of Yale University asserted that given the interactive complexity and tight coupling characteristics of certain systems, 'normal accidents' where multiple and unexpected interactions of failures with humans were inevitable as complex and tightly interconnected technologies are by their very nature, unsafe.

The implications of normal accidents are:

- Operators are confronted by unexpected and mysterious interactions among failures (so anticipation is of limited use with complex and interactive technologies).
- Great events have small beginnings.
- Organisations and management play a major role in causing (and preventing) accidents and failures.
- Fixes, as well as safety devices, add to the inherent complexity and thereby, to the likelihood of accidents.

Components can thus affect each other unexpectedly and are also capable of spreading problems. Adding safety components may increase the range and scope of potential interactions and therefore the number of potential ways for something to go wrong. In other words, safety interventions can redistribute the burden of risk rather than reduce it. This redistribution may be unpredictable and uncontrollable; suggesting that shifting risks may be more dangerous than tolerating them.

[1] Wiener & Nagel (1988) and Owen (2001)

6. A question of responsibility?

The deployment of new technologies has always invoked questions regarding their potential harm and their impacts on humans, civil society and the wider environment. However the rapid growth of automation, artificial intelligence and machine learning is raising important new questions about the moral responsibilities associated with using such technologies.

Indeed, if a machine is no longer a tool or instrument used by a human agent, whose morals and conventions does it follow? Moreover, given the autonomy invested in such systems, how are the ethics and rules programed? Ultimately, given the profound uncertainty, complexity and interconnectedness, the inability to consider all potential future outcomes, and, the disruptive potential of inconsistencies as exhibited in the Airbus crash, who bears the ultimate responsibility for the impacts of such interaction between technology and society?

Just because we can design all types of programs does not mean that we should Rapid development of automation raises questions about the safety of the artefacts being delivered. Given the societal implications of new technologies, it is important to ensure that the impacts of change are understood. Driverless cars, drones and other autonomous creations are likely to transform and revolutionise roads, delivery systems and most other aspects they interact with.

It is therefore important to promote responsibility for the autonomous nature of new technology. Robots are only as reliable as the system that built them. Developers play a key part in shaping the new technological revolution, but need to be held to account for their role in delivering it. Ultimately, when it comes to execution, systems will follow the programmed instructions. The tricky part is to make the code comprehensive enough to cover all eventualities and to exhaustively test it to ensure it is safe, as consumers need to know that it is reliable and trustworthy.

It is difficult for policy makers and consumers to keep up with rapid developments in autonomous technology and robotics, yet it is crucial to ensure that consumers are protected. Whether we recognise it or not, assumptions related to risk and uncertainty are embedded into all the artefacts that we develop. If we send our children to school in a driverless taxi, we would like re-assurance that it will minimise risks on the journey. Does that mean that before entering a junction or joining a roundabout, the driverless car will wait forever, for risk-free entry?

Issac Asimov explicitly commanded robots not to do harm, but that act requires recognition of harm and consequences. Indeed, why assume that robots will seek to inflict damage in the first place? The trolley dilemmas show that we need to understand the worldview of developers and determine if they are looking at situations as consequentialists or categorical moralists. Responsibility ultimately lies with the designers and promoters of new technology.

While responsibility entails owning up to acts, effects and consequences, one can discern different types of responsibility:

- ❖ **Causal Responsibility:** associated with bringing something about either directly or indirectly (e.g. by ordering someone else).
- ❖ **Legal Responsibility**: associated with fulfilling the requirements for accountability under the law.
- ❖ **Moral Responsibility**: associated with having a moral obligation or with the fulfilment of the criteria for deserving blame or praise for a morally significant act, or omission, and the resulting consequences.
- ❖ **Role Responsibility**: associated with duties that are attached to particular professional, or societal, (or even biological) roles. Failure to fulfil such duties can expose the role-holder to censure, which can be moral, legal or constitutional.

Moral responsibility normally assumes some degree of causal responsibility. Therefore a professional can be held morally responsible for failing to act. When we take control from human experts, such as pilots, and offer it to machines, we re-design the responsibility equation. Developers would potentially bear causal, legal and moral responsibility for events. They may also be held accountable under the obligations of role responsibility. Indeed, as we engage with new technologies, apportioning responsibilities may need to become a key activity.

7. Designing for trust

Can we trust new technologies? Trust is often established on the basis of the reliability of a system. This is a problem for new systems with no known track record. It is also a problem when safety features are added and are expected to operate in concert with existing components. Moreover, the addition of new safety features can also impact the reliability of a system, by introducing new modes of failure. With safety viewed as an emergent property of an entire system, the relationship between the different properties can be depicted in Figure 1. Trust in the safety of a system would require fundamentally different tests to establishing trust in the reliability of the components.

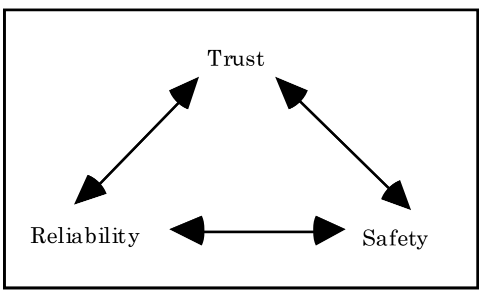

Figure 1. Reliability, Safety and Trust

Some of the new autonomous technology seems to require a maxim of 'trust until proven otherwise'; however, this is a dangerous position. System safety depends on the interaction between components (rather than on past history) as well as considering the place of humans within the system. Trust in a system needs to be built on the basis of considering all aspects and their relationship, prior to the release of the ultimate technology.

Trust is not simply a function between the client and the product system. It is a complex mechanism that involves the developers and that should include trade-offs and understanding of the different moral reasoning systems to balance the different concerns and account for different ways of considering consequences, impacts, and the range of permitted and forbidden operations and interactions (see Figure 2).

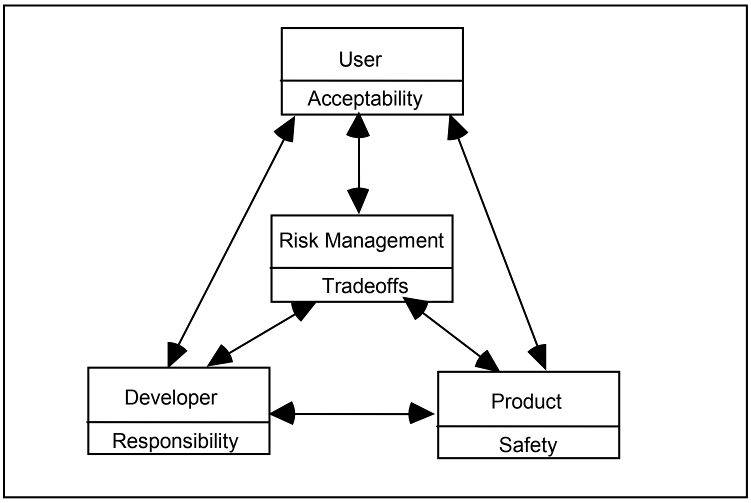

Figure 2. The Dynamic Balance of Trust

Trust applies to the relationships between the user and the developer, the user and the product and, the developer and the product. Developers have a direct influence over the safety level of the system. Users and operators can trust developers as easily as they trust systems. Past record and reputation is likely to influence the way developers are perceived. Developers discharge their responsibility to their client (hopefully courting additional trust) by developing the product. Risk management is central to balancing and trading-off acceptability, responsibility and safety levels and hopefully moral reasoning, thereby enabling direct relationships between risk and acceptability, risk and responsibility and, risk and safety. For example, rather than deal with absolute safety, the user can view safety as a measure of the acceptability of some degree of risk.

Trust and acceptability are also coupled to the ability to control systems, hazards and risk levels. Systems delivered to users should combine the elements of trust and acceptability with an agreed level of control. Wresting control from operators and users should imply attaining their full trust in the system. The model thus offers a new way of reasoning about the adequacy of designed systems.

8. Trust reprised.

Trust is fragile: It is created slowly, but can be destroyed in an instant. Trust builds up over time as a result of complex multidimensional interactions.

To establish trust, there is a need for designers to take their responsibility role into account. Ultimately, the responsibility for a developed system lies with the developer. It is useful to refer to Hammurabi, King of Babylon, who recognised the perils of design some 3,570 years ago and enacted a building code that clarified the 'responsibilities' of designers:

"If a builder has built a house for a man and his work is not strong, and if the house he has built falls and kills the householder, that builder shall be slain."

While, the sentiment may seem harsh, it useful to apply a personal test to new technology: Would you place your child in the hands of the new technology which you are about to design, sell or commission?

The final word on the topic is reserved for Astronaut Alan B. Shepard, who whilst awaiting blast-off atop the space shuttle Columbia, commented that it was a humbling experience knowing that his fate depended on a vehicle built by the lowest bidder!

As we embark on our own journey into the realms of uncharted technology that will transform our future, we could humbly join Shepard in reflecting on the potential impact of a scary new technology, our limited knowledge of its working, and the trust that we must engender in its ability to do good (or at least to endeavour to do the least harm).

References

Asimov, Issac (1950). *I Robot*. Gnome Press.

Beaty, D. (1995) *The Naked Pilot: The Human Factor in Aircarft Accidents,* Airlife, London.

Owen, D. (2001) *Air Accident Investigation*, Patrick Stephens, Yeovil.

Perrow, C. (1984) *Normal Accidents, Living with High-risk Technologies,* Basic Books, New York.

Shelley, Mary Wollstonecraft (1818). *Frankenstein; Or the Modern Prometheus*. Lackington, Hughes, Harding.

Weizenbaum, J. (1976) *Computer Power and Human Reason,* W. H. Freeman, New York.

Wiener, E. and Nagel, D. (Eds.) (1988) *Human Factors in Aviation,* Academic Press.

Woods, D. D. (1991) In *Human-Computer Interaction and Complex Systems* (Eds, Weir, G. R. and Alty, J. L.) Academic Press, London.

Further reading

Barrat, James (2015) *Our Final Invention: Artificial Intelligence and the End of the Human Era*. St. Martin's Griffin

Brockman, John (2015). *What to Think About Machines that Think*. New York: Harper.

Carr, Nicholas (2015). *The Glass Cage: Who Needs Humans Anyway?* London: vintage.

Chace, Calum, (2014) *Pandora's Brain*, Three Cs.

Chace, Calum, (2015) *Surviving AI*, Three Cs.

Lin, Patrick, Abney, Keith and Bekey, George (2014) *Robot Ethics: The Ethical and Social Implications of Robotics*. Cambridge: MIT Press.

Wallach, Wendell and Allen Colin (2009) *Moral Machines: Teaching Robots Right from Wrong*, Oxford: Oxford University Press.