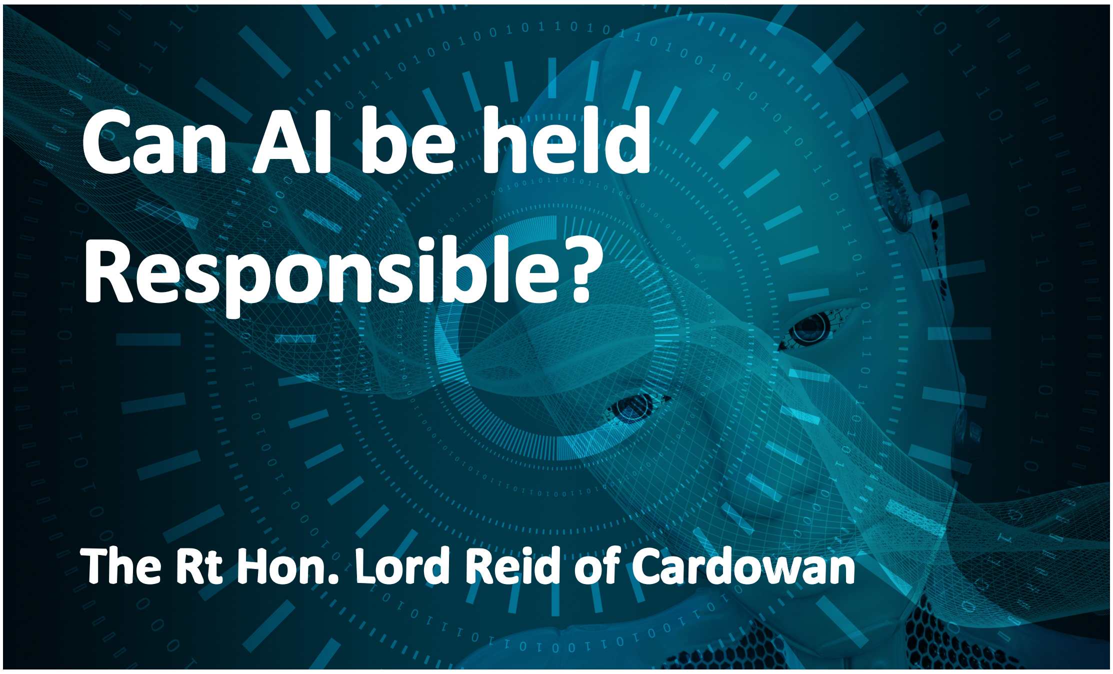
TRANSCRIPT

Institute for Strategy
Resilience & Security
University College London

Bird & Bird | Tech Law Day | 9 Oct 2018



Can AI be held Responsible?

The Rt Hon. Lord Reid of Cardowan

As we stand on the edge of a new world looking forward, and project out the evolution of technology, there is no question that “AI” will play a huge role in our future. In the space of a century, machines have evolved from simple, deterministic, mechanical devices to embody sophisticated intelligence.

Although technology is evolving at an extraordinary rate, it is sobering to put into perspective the power of the human brain, as a result of several billion years of evolution in forming neural circuitry. Five years ago, a team of scientists at the Institute of Neuroscience and Medicine at Jülich in Germany, attempted to simulate just 1% of the human brain -- a network of 1.7 billion nerve cells connected by more than 10 trillion synapses. To simulate just 1 second of this 1% brain activity required no less than 40 minutes of processing time of Japan’s “K” computer, one of the world’s largest supercomputers, equivalent to 250,000 PCs.

So think of that: the equivalent of a quarter of a million PCs was still a quarter of a million times slower and less powerful than the human brain, or to put it another way, a single human brain, consuming less than 20 watts of power is 60 billion times more powerful than a PC. If you truly want to make a “human” decision today using a computer, that decision process would still need to use more than all of the world’s computing devices put together.

We need to distinguish our current stage of machine learning, from what is meant by true artificial intelligence. Today, machines are being taught to display intelligence or behave intelligently -- that doesn’t mean that they are yet intelligent or sentient. Machine learning is ever more functionally powerful and effective at emulating human-like reasoning and emotion. These increasing capabilities are driving us to trust machines more, both because they are demonstrably equal to, or better at controlling complex functions and systems, for example an autopilot in an aeroplane, or because their emulated behaviour appears human-like Alexa or Siri. The systems that we are used to ceding control to, while functionally superior, are however largely deterministic or operate within bounded parameters. What happens when we cede control to non-deterministic and self-evolving systems? What happens when these systems control vital functions?

We have defined AI in our self-image with the Turing test: an entity that is indistinguishable from human intelligence. Machine learning is increasingly being used to design human-like presences. Devices that respond with human voices, call centre operatives who are not real yet behave like humans. In the future, driven by commercial, military and scientific competitive motivations, AI will be designed and will further self-evolve to have both more sophisticated machine intelligence functions and more human-like interfaces.

Machine intelligence will provide greater functional or predictive value, while more human-like interfaces will improve the perception of empathy and therefore value-extraction from interactions with humans. The ability to appear human-like, based on the ability to recognise facial emotions accurately and to conduct sophisticated, responsive conversations in any language, about any topic, for example, will allow organisations to project human-like responsibility from what are actually software agents.

This is creating an asymmetry, because when something goes wrong, who takes responsibility for sorting out the problem? It becomes increasingly easy and desirable for every party in the value-chain to absolve blame.

The more convincingly “human-like”, the more we may be tempted to believe that the underlying motivations of AI are human. However, despite human-like appearance and superior functional control, the underlying architecture of AI will remain “alien” -- the motivations, stimuli and the neural circuits that result in its outputs are not human. Machines may take actions that are functionally “correct”, but they are not doing so out of concern or empathy. If an autonomous vehicle swerves to avoid a human, it is not acting because it understands the value of life, or empathises with the pain that it might cause, or a sense of consequence of its actions. Machines and humans are both prone to error and deception, and under certain conditions may behave very differently and unexpectedly. Driven by conflicting directives and goal functions set to serve humans, they are almost certain to become deceptive, in the same way that humans do – Remember HAL in 2001: A Space Odyssey?

Many forms of responsibility exist: legal, ethical, and moral. Today in law we make a distinction that a human may be responsible, while a machine or an animal may not be. We seek to distinguish fault from genuine accident, where no one is responsible;

or, where fault has been established, to assign individual or corporate culpability. However, in cases such as a dog biting a human or an animal escaping from a zoo, the animal may be put down because it is a danger, despite having no legal, moral or ethical culpability.

As humans we have law-based moral obligations as part of our social contract within a civilised society; we have promise-based moral obligations as part of contracts that we form with others; and we have societal moral principles that are the core of what we regard as ethics, whether derived from rational reason or religion. These are not merely extremely complex to encode within the intelligence of an AI entity, but despite increasing intelligence, its underlying logic will never be human because its motivations, inputs and neural circuitry are fundamentally different. AI may ultimately achieve sufficient intelligence to exhibit or to emulate empathy.

But this is a key distinction. When a railway announcement says, “We are sorry to announce…” the computer voice is not “sorry”, neither is the Company operating the service – neither can be, because they cannot experience emotions or remorse. However, the Company perceives sees value in appeasing its customers by offering an apology. Thus, AI entities will likely be emulating emotion, long before they are feeling it.

Once they truly empathetic and compassionate – capable of feeling suffering and happiness in a manner that is equal to or exceeds human – then they may equally need to be given the same rights as a human.

However, a question remains - that if you are not actually capable of experiencing pain and suffering, you cannot truly be empathetic. The threat of punishment or imprisonment cannot have any rational meaning to a machine entity, whose perception of time, if it were to exist, would be utterly different from our own. The “death penalty” would have no meaning to an AI entity if it viewed its reincarnation (being “rebooted”) as a positive benefit – a bit like the motivations of suicide bombers who believe in martyrdom.

Even were machine intelligence to exceed human intelligence -- likely as we project out the rate of technological advance -- we must distinguish between being better at carrying out a set of tasks and human responsibility. Intelligence is not the sole determinant of responsibility within human society – the “age of responsibility” that distinguishes a minor from an adult, is based on the inability of children to make good decisions, being too immature to understand the consequences of, or consent to, certain behaviour. A machine also cannot be punished or be incarcerated in any meaningful sense, although it might be rehabilitated through reprogramming. The notion of consequence of actions therefore has little meaning to a machine. If a machine apologises or serves a prison sentence, or is put in solitary confinement, has it been punished?

We would argue that machine intelligence therefore needs to be subordinate in responsibility to a human controller, and cannot therefore be legally responsible as an adult human, although it might have the legal status of a corporation or, in the future as a minor – i.e. intelligent, but below the age of criminal responsibility.

Machine intelligence, unlike a physical device, may also be spread across multiple devices and locations. We need a process to ensure that the people responsible for its design and use remain accountable despite the diffuse nature. GDPR was designed for passive "data" and ultimately we may need a GDPR-equivalent for active machine learning systems to link their function to a human controller, and ensure that organisations and individuals have protective and proportionate processes in place.

Responsible humans are aware of ethical, moral and legal obligations – we feel empathy towards fellow humans. We feel a responsibility to our children, employees and society. Those who don't are called psychopaths or sociopaths. How do we encode concepts such as "the public good" or "volunteering" in the goal-functions of machines? These are perhaps far harder challenges than the demonstration of functional machine intelligence.

Until AI is demonstrably superior to humans in its ability to make moral, ethical and social decisions, this lack of emotional parity with humans could be seen at best as immaturity, or at worst as sociopathic or even psychopathic. Psychopaths, and to a degree, sociopaths, show a lack of emotion, especially the social emotions, such as shame, guilt, and embarrassment. AI will be designed and will evolve to appear superficially to be human but will not exhibit any of the underlying empathetic emotions of a human, except those designed to elicit the desired responses of those from whom they are intended to extract value.

AI will require continual psychological evaluation to ensure that it is acting in the desired manner, as well as human supervision and an accountability mechanism to ensure that someone - a human - takes responsibility for its actions.

One major issue is that today most software is explicitly sold with a disclaimer of fitness for purpose. In practice it is virtually impossible to look back and ask, by whom, against what specification, when and why was code generated, tested or deployed.

In industries that are critical to modern society – the notion of audit exists. When we read the annual report of a PLC, it is possible to place some degree of trust in it because the CFO, accountant and auditor are taking professional responsibility for the output. The process doesn't ensure that the annual report is "perfect", but it does mean that if something subsequently turns out to be amiss, the issues can be traced back to the persons who signed it off.

Equally within the pharmaceutical industry, the audit chain enables regulators to oversee a large, complex and diverse industry. All of the players are acutely aware of the consequences of regulatory infringement. In the event of a bad drug batch, the regulator is able to determine where the manufacture took place.

In Construction, when a tragedy happens we are able to trace back the building materials used in construction. The point being that foreknowledge of accountability drives greater quality assurance.

In the event of a problem with software who is responsible: Its human owner, the Company that supplied the software, the programmer, or the CEO of the company that supplied it?

At ISRS we refer to the concept of a clear chain of responsibility, linking an audit of their specifications, code, testing and function to responsible individuals, as trustable, a topic we are working on today. In practise this is very hard to determine. If an autonomous vehicle were involved in a fatal accident, what information would be available about the crash? We would surely expect self-driving cars to contain a "black box" much like the aerospace equivalent to provide an audit trail of the actions that led up to the incident.

Will machines have the legal status of humans, physical objects, animals or corporate entities? Physical objects that are controlled by a person are regarded as extensions of that person – the person is to blame, not the car or the gun. If the car or gun were to malfunction, the manufacturers may be responsible.

Animals were recognised as property at a time when leading philosophers believed that God had given humans dominion over all animals. It was believed that they did not have any moral standing because they lacked rationality and autonomy. Today we believe that the interests of humans and animals should receive more closely equal moral consideration because both have the ability to suffer, feel pain and experience enjoyment.

Certainly, like animals, AI is distinct from other forms of property, such as tables, land and intellectual property. However unlike inanimate objects, AI has the capacity for independent action. In future it may not only engage in intelligent thinking but also have the capacity to experience and display suffering. By categorising AI as property, the law objectifies it - animals cannot be bearers of legal rights. Legal personality, or standing, is a precondition for having and enforcing legal rights. Once it is accepted that AI is sentient and therefore has morally relevant interests, it becomes difficult to justify the treatment of such beings as mere objects.

In the case of an autonomous machine, the lines blur:

- If AI becomes functionally indistinguishable from a human, might it need to be treated with the legal status of a "child"?

- If it were to be given responsibility, would it also have be entitled to corresponding rights?

- Will we need to prevent the infliction of forms of harm on AI?

- Will we need to limit their exploitation? Will we need to ensure that they have adequate power supplies?

At some point AI may evolve sufficient intelligence and display sufficiently sophisticated cognitive abilities that we judge it to be superior in ethical and moral decision making to individual humans or even collective humans. The notion of “singularity” is of artificial general intelligence that self-evolves exponentially in sophistication, ultimately to have an intelligence that exceeds our own, or even the collective intelligence of humanity.

At that point, logically, we may have no ability to judge AI’s truth or wisdom, because we will be the inferior intelligence. However, the problem will remain that based on a different cognitive architecture, we may never be entirely sure whether it is truly acting in our best interests, or whether it is merely trying to game or control us, as an inferior species, for our own good or perhaps its own.

A final point, as we gaze into the future is to remind ourselves that individual human responsibility is not the highest form of responsibility within human society. We acknowledge that some level of decision taking requires collective wisdom, and the need for government by a collective. We understand that better decisions are made by boards within companies, the need for governments made up of many individuals, and the importance of the exercising the will of the people in referenda, rather than autonomous decision by an individual, acting as a benevolent dictatorship. This is because an individual, however brilliantly intelligent, may not necessarily act wisely or in everyone’s interests.

As we shift to an age where software **is** the infrastructure, where all systems are interconnected and where machine learning is generating non-deterministic systems, we move towards the notion that we can only really evaluate whether its outputs are empirically those that were desired, and that AI therefore needs a responsible human "parent". And - whoever that “parent” might be - it will require a “trustable” process to introduce auditability, accountability and, ultimately, responsibility.

ISRS makes every effort to ensure the accuracy of written transcripts, but due to the nature of transcribing recorded material, this transcript may contain errors or incomplete content. It may be imprecise or capable of different interpretations. As such, any information or opinion expressed in the transcript should be cross-referenced to other published, verified material before relying on it. If you wish to take direct quotes from the transcript, please use the webcast of this briefing or contact us to confirm their accuracy. This transcript may contain forward-looking statements that involve substantial risks and uncertainties. You can identify these statements by the use of forward-looking terminology such as “may,” “will,” “should,” “expect,” “anticipate,” “project,” “estimate,” “intend,” “continue” or “believe” or the negatives thereof or other variations thereon or comparable terminology. Any forward-looking statement made in this transcript speaks only as of the date on which it was made.

TRANSCRIPT

ABOUT THE INSTITUTE FOR STRATEGY, RESILIENCE & SECURITY (ISRS) AT UCL

The Institute for Strategy Resilience & Security (ISRS) (https://www.isrs.org.uk) at UCL serves as a pioneer and forum for next generation thinking. Founded by the Rt Hon. Lord Reid of Cardowan, ISRS provides analysis and assessment of the major issues of resilience with respect to national and global infrastructure and the ability of governments, regulators and businesses to respond to them. The Institute advises industry and the public sector on the persistent challenges to their agility, stamina and capacity for strategic decision making, so as to better face existential threats and disruptive innovation that are not addressed by conventional strategy and forecasting.

PRESS CONTACT INFORMATION

Institute for Strategy, Resilience & Security (ISRS)

University College London

Gower Street

London

WC1E 6BT

Tel.: +44 (0) 207 193 0060

E-mail: media@isrs.org.uk