My Robot: AI, Ethics and the Unthinkable

Today's robots already break Asimov's laws of robotics. As the age of artificial intelligence dawns, how will we govern the behaviour of machines when we can no longer truly understand their code or even know how they may respond as they evolve through machine learning? Who will set the standards by which both human and machine societies will live?

Cassandra Kelly
Founder and
Senior Advisor
CassandraLKelly

Nigel Lake
Founder and
Executive Chair
Nigel_Lake

#AI #Governance #Robots #Tech

Pottinger
PREVIEWS

My Robot: AI, Ethics and the Unthinkable

A robust global framework for the supervision of machines that employ any form of AI or machine learning should be implemented urgently and must be effectively policed.

Manufacturers must retain first-line responsibility for the behaviour of all machines they produce

All robots must have **failsafe systems**, **data capture** and storage, and robust **kill switches**

Universal basic laws for the governance of robots must be agreed globally and carefully policed

"The rise of powerful AI will be either the best, or the worst thing, ever to happen to humanity. We do not yet know which." **Stephen Hawking, 2016**

Artificial intelligence has many connotations. For us, the most important is the shift from machines executing predictable, pre-programmed tasks to computers taking responsibility for decision-making. This is the Robot Revolution – a fourth great societal leap following the agricultural, industrial and technological revolutions.

Most commentators agree that AI will enable unprecedented human advancement and also create significant new risks for individuals and society at large. For now, computers can still only tackle narrowly-defined tasks. Soon, however, the vision or spectre of general artificial intelligence will appear. One thing is certain: too much is at stake to watch from the technological side-lines.

In his seminal 1940s work *I, Robot*, Isaac Asimov set out three fundamental laws of robotics, which every robot must obey without fail. To this day, these are an excellent foundation for considering how to ensure robots operate within appropriate societal norms.

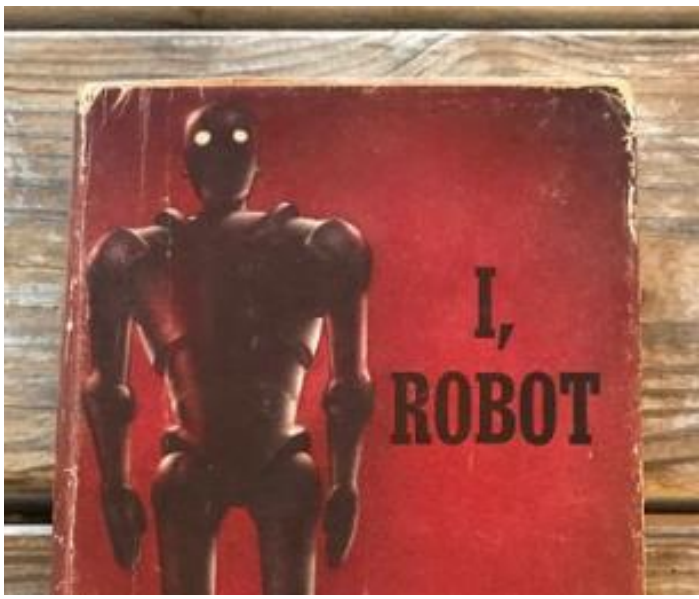
1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2. A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.
3. A robot must protect its own existence as long as such protection does not conflict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

Yet, some robots already break these laws. To avoid an accident, self-driving cars may make decisions that injure other road users (breaking the first law). Flying robots (ie semi-autonomous drones) are regularly used to take human life (breaking the second law).

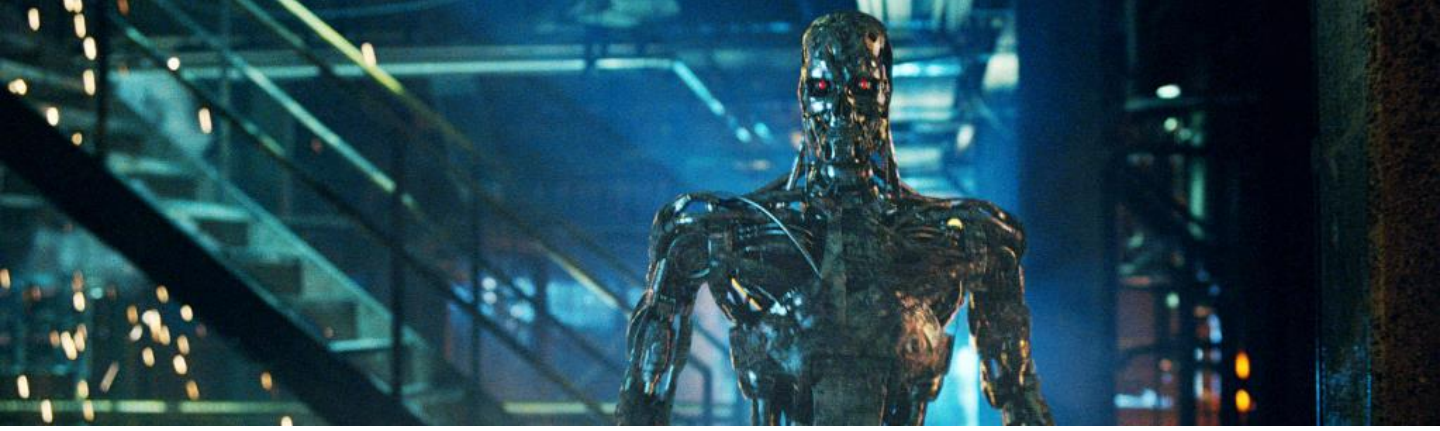
Importantly, Asimov's laws focus on human safety, followed closely by the need to ensure that humans retain control over the machines. As technology evolves, we will need governance frameworks that protect human rights, preserve economic and social wellbeing, and ensure that machines do not become a law unto themselves. Though this detail is important, we must not let it obscure the primary importance of safety and control. We must therefore address the overarching ethics of the behaviour of artificial intelligence systems separately from the broader discussion related to how the benefits of AI are shared with society.

Whose ethics is it anyway?

As the use of AI expands, robots' decisions may not be predictable by humans. This may occur because so much data is involved that humans would simply not have time to comprehend it all, or because the complexity of the problem being solved is literally beyond human understanding. Moreover, thanks to machine-learning algorithms, any robot's thinking may evolve. As a result, two identical robots that are released into the wild and exposed to different data sets may learn to adopt different approaches to the same task as time goes by, as each amasses new and different experience.



Pottinger



To address this challenge, robopsychology must now move from Asimov's science fiction into the real world. We must ensure we understand how any robot's behaviour is likely to evolve. And we must have absolute clarity regarding who will be held accountable for the actions of a wayward machine.

We propose that the original manufacturer of the robot should have first-line accountability for the behaviour of its machines. To be clear, we are not seeking for manufacturers to have ultimate responsibility for the behaviour of their machines, so long as it can be proven that another party is responsible. What we are seeking is that the burden for inventing and implementing appropriate command, audit and control systems lies with the party that has brought the machine into the world.

Although this may raise significant legal and commercial issues for the organisations concerned, we note that progressive businesses are already adopting this type of mindset, recognising that without the right controls in place they may lose their social licence to operate. One example is drone manufacturers, whose software already enables permanent or temporary 'no fly' zones to be implemented around sensitive locations.

How can we police the behaviour of machines?

Next, we must define clearly what society expects of robots and their owners to ensure human safety and absolute control over the machines. In this context, we note that robots created to address specific problems will be much easier to manage and control than those which are equipped with more general artificial intelligence and hence are able to tackle a much wider range of tasks.

We see three elements to these protections that every robot manufacturer should embrace:

1. Failsafe systems to prevent the occurrence of events which, if committed by a human, would be crimes, misdemeanours, or other breaches of relevant laws or regulations;
2. The ability to capture and retain data in a manner that is sufficient to provide an audit trail to demonstrate why particular decisions were made, and hence to provide objective evidence as to who should take responsibility; and
3. Robust kill switches that would ensure any robot could be shut down remotely if this was required to preserve human safety and/or control over the machine in question.

Though these elements are simple to describe, they will be challenging to implement, especially given the diversity of organisations that are involved in creating robots.

Who will make and police the robot law?

Finally, as should already be obvious, a 'free market' or 'self-regulated' solution for AI ethics is entirely inappropriate. There are many complex issues and risks at stake, and without proper governance the short-term interests of robot sellers and robot users will swamp the longer-term interests of society.

Accordingly, we believe it is imperative that governments and intergovernmental agencies co-operate closely to adapt basic universal laws to address the ethical issues raised by artificial intelligence, much like the fundamental human rights that have been established and agreed over recent decades.

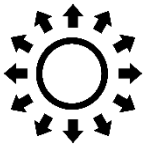
Given the pace of development, action is urgent. As Dave Waters commented: "If Elon Musk is wrong about artificial intelligence and we regulate it – who cares? If he is right about AI and we don't regulate it, we will all care.".

For further information, please visit Pottinger.com/Previews

Consulting | Investment Banking | Implementation

Strategy and public policy

M&A, JVs and investments



Innovation and digitalisation

Risk and data analytics

People and leadership

New York | London | Sydney

Pottinger PREVIEWS