Can you trust a robot?

Accountability, compliance and ethics in the age of smart machines

We have never known a technology that can more greatly benefit all of society than artificial intelligence.

The ability of AI to transform vast amounts of complex, ambiguous information into insight has the potential to reveal long-held secrets and help solve some of the world's most enduring problems. It can be used to help discover and treat disease, predict the weather and manage the global economy. AI is an undeniably powerful tool. But like all powerful tools, great care must be taken over how it is used.

Great expectations

To reap the full benefits of AI, we first need to trust it. The right level of trust will be earned through repeated experience, in the same way we learn to trust that a car will stop when we hit the brakes. Put simply, we trust things that behave as we expect them to.

But trust will also require a system of best practices that can help guide the safe and ethical management of AI systems, including:

- Alignment with social norms and values
- Algorithmic responsibility
- Compliance with existing legislation and policy
- Assurance of the integrity of the data, algorithms and systems
- Protection of privacy and personal information

A familiar story

We have been here before. In fact, society has repeatedly developed and adjusted to advanced technologies – from the ancient discoveries of the wheel and simple tools, to the invention of the steam engine and the Internet itself. Each has led to significant changes in the way we live and work. Each has supported major improvements in our quality of life. And each has been accompanied by unexpected, and in some cases unwanted, consequences that must be managed.

Most recently, it was the Industrial Revolution that forever changed the course of human history. Artificial intelligence, it would seem, is now forming the foundations of a "second machine age".

We believe that by combining the best qualities of machines today – and in the future – we will come to understand our world better, and make more informed decisions about how we live in it. AI systems will enhance our ability to learn and discover, opening new avenues of thought and action, and pushing us to higher, more meaningful planes of existence.

Not surprisingly, the prospect of this "second machine age" has been accompanied by a significant amount of anxiety. This is neither unusual nor unreasonable. The early days of any new technology are often met with resistance. Change always makes us uncomfortable. And that discomfort is fertile ground for myth and misunderstanding to take root.

A matter of semantics

Perhaps our biggest obstacle in quelling the general anxiety over artificial intelligence is semantic. The term “artificial intelligence” historically refers to systems that attempt to mimic or replicate human thought. This is not an accurate description of the actual science of artificial intelligence, and it implies a false choice between artificial and natural intelligences.

At IBM, we are guided by the term “augmented intelligence” rather than “artificial intelligence”. This vision of “AI” is the critical difference between systems that enhance, improve and scale human expertise, and those that attempt to replicate human intelligence.

Common AI myths

1. They’re taking our jobs

While artificial intelligence will almost certainly redefine work in many industries, it will also lead to net new industries, companies and jobs – many of which are difficult to even conceive at this organisations in the world, indicates that technological advances like AI lead to net job growth.

2. Our information won’t be safe

When it comes to the protection of personal information, many of the same concerns that exist around today’s computer systems also apply to AI. It is true that AI systems will be more capable of uncovering net new information from personal information, and that new insight will need to be protected with the same level of rigour as before. But we think that AI will actually help solve this problem, and be better at protecting privacy through advanced techniques like de-identification and privacy-preserving deep learning.

3. Humans are being replaced

The notion of artificial general intelligence (AGI) – an autonomous, self-aware AI system with all human abilities including consciousness – is an extremely ambitious goal, for which our scientific understanding is in a supremely early phase. We believe that much progress and benefit will come from the practical approach of specialised AI – systems that support tasks in well-defined domains – before AGI can even be contemplated. In the meantime, we are working with our clients, business partners and competitors to put in place best practices for the safe deployment of a range of AI systems.

We must do what is right

As with any tool, physical or digital, there will be instances in which AI can be used unethically. Our job as a technology company and a member of the global community is to ensure, to the best of our ability, that any AI we develop is created in the right way and for the right reasons.

Intent

Every company should have in place guidelines that govern the ethical management of its operations and the conduct of its employees, as well as a governance system that helps ensure compliance. These guidelines should restrict the company from knowingly engaging in business that would be detrimental to society. And those same standards of ethical business conduct should guide the development of AI systems.

Algorithmic responsibility and system assurance

Trust is built upon accountability. And the algorithms that underpin AI systems need to be as transparent, or at least as interpretable, as possible. In other words, they need to be able to explain their behaviour in terms that humans can understand – from how they interpret their input to why they recommend a particular output.

To do this, AI systems should include explanation-based collateral systems. These systems already exist in many advanced analytical applications for industries like healthcare, financial services and law. In these scenarios, data-centric compliance monitoring and auditing systems can visually explain various decision paths and their associated risks, complete with the reasoning and motivations behind the recommendation. And the parameters for these solutions are defined by existing regulatory requirements specific to that industry, such as HIPAA or Basel III.

Embedded values

AI systems should function according to values that are aligned to those of humans, so that they are accepted by our societies and by the environment in which they are intended to function. This is essential not just in autonomous systems, but also in systems based on human-machine collaboration, since value misalignment could preclude or impede effective teamwork.

There are two main approaches to embedding ethical values into AI systems:

- Top-down: recommends coding values in a rigid set of rules that the system must comply with. It has the benefit of tight control, but does not allow for the uncertainty and dynamism AI systems are so adept at processing.
- Bottom-up: relies on machine learning (such as inverse reinforcement learning) to allow AI systems to adopt our values by observing human behaviour in relevant scenarios. But this approach runs the risk of misinterpreting behaviour or learning from skewed data.

We believe that a combination of top-down and bottom-up approaches would be practical, where coded principles and ethical rules can be dynamically adjusted through the observation of human behaviour.

Robustness (verification and validation testing)

Robustness is a measurement of the reliability and predictability of systems. As such, it is a critical requirement of establishing the right level of trust in an AI system. To achieve robustness, all AI systems must be verified, validated and tested, both logically and probabilistically, before they are deployed.

Verification is a technique in computer science to confirm that a system satisfactorily performs the tasks it was designed to perform. Because AI systems operate in partially unknown environments and act upon ambiguous information, new verification

techniques will be required to satisfy this aspect of robustness. Validity is another technique to gauge predictability, and thus confirm that a system does not have unwanted behaviours (and thus consequences). To define those unwanted behaviours, we need to know what is good or bad in a particular situation, referring back to embedded values. Because there is a risk of emergent behaviours with AI, this process must be ongoing and overseen by human beings.

In practice, this takes the form of extending existing practices for requirements management and field testing that are part of today's product management life cycles. The notion of alpha and beta field testing will have to be redefined to incorporate the probabilistic behaviour of AI systems.

A collective effort

Defining and embedding ethical guidelines is only half the battle. Helping to maintain compliance with those guidelines is a longer-term prospect. And this will be a collective responsibility – shared by the technology companies developing artificial intelligence, the industries applying them and the regulatory agencies that oversee safe and fair business practices.

Each member of this community is obliged to make their efforts transparent and collaborative. The future of this vital technology, and most importantly the benefit it can bring to all of humanity, depends on it.

Join us

We must all keep this conversation going. Because there is too much to gain to let myth and misunderstanding steer us off our course. And while we don't have all the answers yet, we're confident that together we can address the concerns of the few for the benefit of the many.

"Nothing in life is to be feared, it is only to be understood. Now is the time to understand more, so that we may fear less."

Marie Curie

The work of understanding our responsibilities in developing and deploying safe and ethical AI systems is ongoing. And the development of trust will come through use over time, just as trust was built with all technologies that preceded AI, and all that will follow it.

As the technology develops and matures, we encourage other technology companies, as well as experts of many other scientific disciplines, to join us in the study and development of robust, dependable and trustworthy AI applications. Artificial intelligence is not without its risks. But we believe the risks are manageable. And that the far greater risk would be to stifle or otherwise inhibit the development of a technology with the potential to greatly improve the quality of life around the world.

Got you thinking?

Read more about advances in AI and how IBM is driving ethical change.