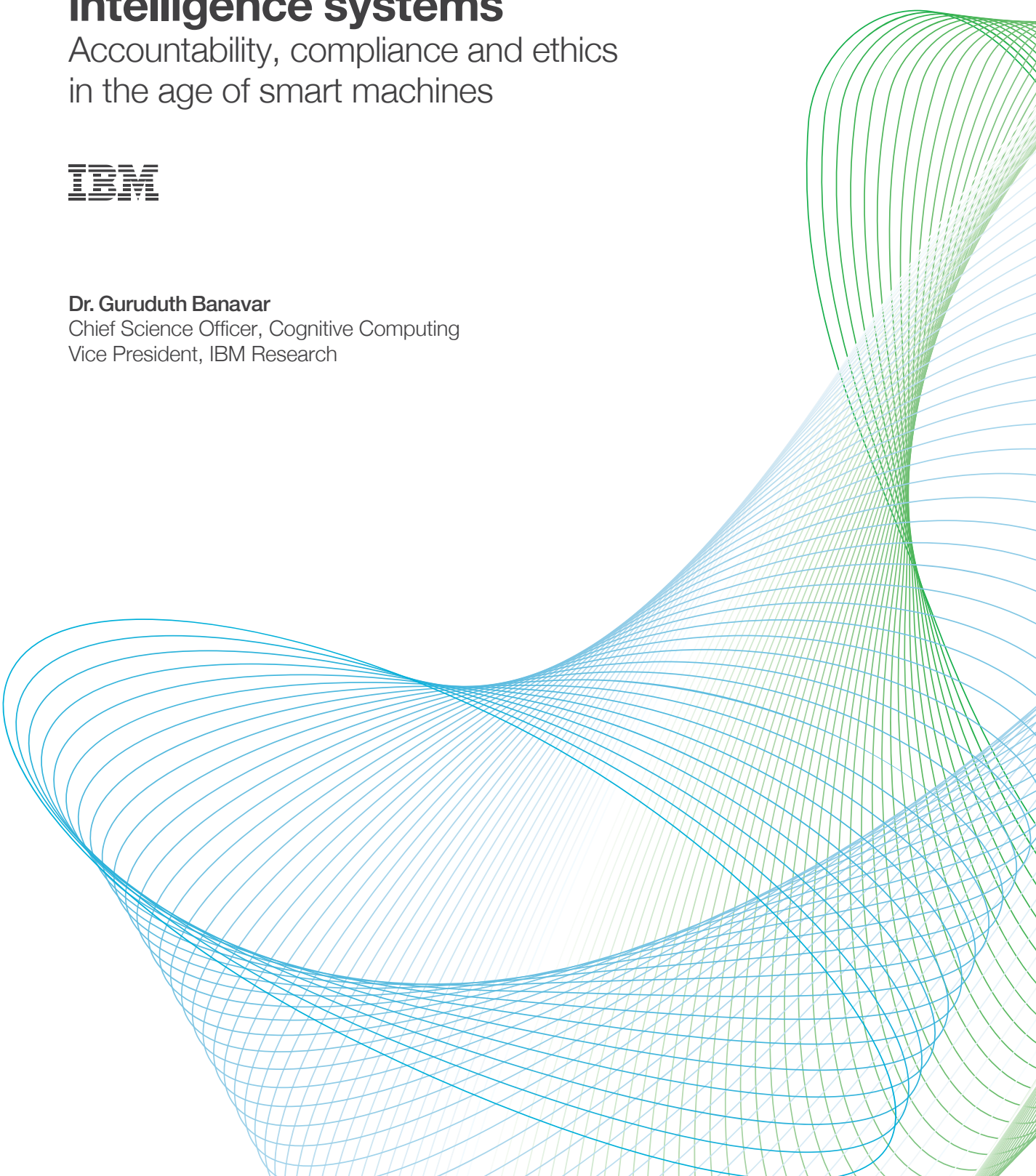
Learning to trust artificial intelligence systems

Accountability, compliance and ethics in the age of smart machines

Dr. Guruduth Banavar
Chief Science Officer, Cognitive Computing
Vice President, IBM Research



For more than 100 years, we at IBM have been in the business of building machines designed to help improve the effectiveness and efficiency of people. And we've made measurable improvements to many of the systems that facilitate life on this planet. But we've never known a technology that can have a greater benefit to all of society than artificial intelligence.

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.

The ability of AI systems 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. AI systems can potentially be used to help discover insights to treat disease, predict the weather, and manage the global economy. It is an undeniably powerful tool. And like all powerful tools, great care must be taken in its development and deployment.

To reap the societal benefits of AI systems, we will 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 an ATM will register a deposit, or that an automobile will stop when the brake is applied. 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; and protection of privacy and personal information.

We consider this paper to be part of the global conversation on the need for safe, ethical and socially beneficial management of AI systems. To facilitate this dialogue, we are in the process of building an active community of thoughtful, informed thinkers that can evolve the ideas herein. Because there is too much to gain from AI systems 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 to the benefit of many.

Artificial intelligence: Something old, something new

We have been here before. In fact, society has repeatedly developed, deployed and adjusted to advanced technologies throughout the course of history, from the ancient discoveries of fire, the wheel and simple tools, to the inventions of the steam engine, the printing press and the Internet itself. Each has led to significant changes in the way we live and work. Each has supported major improvements in 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. Powerful machines allowed us to make and move things that were previously inconceivable. It ushered in an unprecedented era of production and progress that transformed every life and every industry in every country. But it also brought with it major upheaval in labor markets and unacceptable environmental impacts, both of which we continue to address in various ways, from economic and education policies that support the development of new jobs, to global coalitions on climate change.

In their book *The Second Machine Age*, MIT's Erik Brynjolfsson and Andrew McAfee write that they believe the Industrial Revolution was the first machine age, and that artificial intelligence is the foundation of a second machine age, in which "computers and other digital advances are doing for mental power – the ability to use our brains to understand and shape our environments – what the steam engine and its descendants did for muscle power."

We agree with this assessment. In fact, we believe that by combining the best qualities of machines — data analytics, procedural logic, reasoning and sense-making — with uniquely human qualities — value judgment, empathy and esthetics — we will come to understand our world better, and make more informed decisions about how we live in it. We will be able to peer into the digital dark matter — the vast, unexplored universe of unstructured data. And we believe that 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 — especially one that holds the potential for great change — are often met with resistance. Change, technological or otherwise, always makes us uncomfortable. And that discomfort is fertile ground for myth and misunderstanding to take root.

As mentioned earlier, 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.

That is why IBM and others have chosen to use different language to describe our work in this field. We feel that "cognitive computing" or "augmented intelligence" — which describes systems designed to augment human thought, not replicate it — are more representative of our approach. There is little commercial or societal imperative for creating "artificial intelligence." But there's always a place for tools that can enhance our intelligence, support better decision-making and extend human expertise.

A system of trust

We understand that some see the potential for abuse and misuse of artificial intelligence. 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 AI we develop is developed the right way and for the right reasons. These guidelines comprise a system which, when adhered to, can help engender trust among the developers, users and beneficiaries of artificial intelligence. This system includes several aspects.

Intent

Every company should have in place guidelines that govern the ethical management of its operations as well as the conduct of its employees, and 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, at IBM or anywhere else.

The business intent and the requirements of any AI system should be clearly defined and ethically acceptable before development work begins. A specific governance practice is to incorporate the

assessment of an ethics advisor as part of the product management process. A second practice is to perform extensive field testing for ethics-related design issues before a product or service is deployed widely. A third practice is to put in place a mechanism for continuous user feedback specifically focused on potential ethics-related issues.

As an industry, we must pay close attention to the track record of government and non-government buyers, and adhere closely (as we do today) to national trade policies and relevant U.N. mandates.

Algorithmic responsibility and system assurance

Trust is built upon accountability. As such, the algorithms that underpin AI systems need to be as transparent, or at least interpretable, as possible. In other words, they need to be able to explain their behavior in terms that humans can understand — from how they interpreted their input to why they recommended a particular output.

To do this, we recommend all 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.

One of the primary reasons for including algorithmic accountability in any AI system is to manage the potential for bias in the decision-making process. This is an important and valid concern among those familiar with AI. Bias can be introduced both in the data sets that are used to train an AI system, and by the algorithms that process that data. But we believe that the biases of AI systems can not only be

Myth and misunderstanding

In the interests of grounding our discussion of the ethics of artificial intelligence in facts, we'd like to briefly address a few of the most common myths that are coloring the public discourse.

First, 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 early stage. In fact, study after study, from the most respected economic scholars and research organizations in the world, indicates that technological advances like AI lead to net job growth. Perhaps the Organisation for Economic Cooperation and Development (OECD) said it most unambiguously: "Historically, the income generating effects of new technologies have proved more powerful than the labor-displacing effects: technological progress has been accompanied not only by higher output and productivity, but also by higher overall employment."

Second, when it comes to the protection of personal information, many of the same concerns that exist in 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 rigor 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.

And third, the notion of an artificial general intelligence (or 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 specialized AI – i.e., 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 as well as competitors to put in place best practices for safe deployment of a range of AI systems.

managed, but also that AI systems themselves can help eliminate many of the biases that already exist in human decision-making models today.

Furthermore, assuring the integrity of AI systems as a whole is even more important as AI increasingly underlies applications across all industries and aspects of our lives. We must carefully manage the integrity of the data and models underlying our AI systems, as well as the resiliency of algorithms and systems in the face of a wide range of anomalies and threats. Anomalies can be introduced by many factors, ranging from incompleteness to malicious attacks. Techniques and processes to protect, detect, correct and mitigate risks due to anomalies must be integrated end-to-end within our cognitive platform. Careful, risk-mitigating actions of a cognitive system may ultimately provide higher value than highly-tuned but brittle analytics, or highly confident but unreliable decision support.

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.

It is not yet clear *what* values machines should use, and *how* to embed these values into them. Several ethical theories, defined for humans, are being considered (deontic, consequentialist, virtue, etc.) as well as the implications of their use within a machine, in order to find the best way to define and adapt values from humans to machines.

But we do have an idea of *how* to embed ethical values into AI systems. And there are two main approaches to this. First, the so-called "top-down" approach 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. The other approach is often called "bottom-up," and it relies on machine learning (such as inverse reinforcement learning) to allow AI systems to adopt our values by observing human behavior in relevant scenarios. But this approach runs the risk of misinterpreting behavior 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 behavior. In this scenario, we would look to the established ethical norms of existing industries like healthcare and finance to guide the embedded code.

In fact, in industries like healthcare and finance, the relevant professional ethical principles are explicitly encoded and practiced by professionals in the field already. In AI systems designed to help professionals in these domains, these best practices and principles could form the core of the ethics module for such systems. Ethics modules, however, should be constantly adapted to reflect humans' best practices in their everyday profession.

We envision a future in which every AI system will need to have its own ethics module to allow for a fruitful interaction and collaboration with humans in the environments in which it is used. This could be achieved by developing an ethics API that can be adapted to specific professions and real-life scenarios. It would provide the main principles and values the AI systems should base its behavior on, as well as the capability to dynamically adapt them over time to tune them to the real situations that are encountered in that profession or environment. Such a rigorous approach could offer sufficient value alignment without compromising the full problem-solving potential of artificial intelligence.

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, acting 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 behaviors (and thus consequences). To define those unwanted behaviors, 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 behaviors with AI — such as situations where the system combines data from previously separate systems — 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 behavior of AI systems.

IBM's commitment

Defining and embedding ethical guidelines is only half the battle. Helping to maintain compliance to those guidelines is a longer-term prospect. And that will be a collective responsibility, shared by the technology companies that are developing artificial intelligence, the industries that are 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.

For its part, IBM is engaged in several efforts – both internally and externally – to advance our understanding and effecting the ethical development of artificial intelligence. They include:

- The establishment of an internal IBM Cognitive Ethics Board, to discuss, advise and guide the ethical development and deployment of AI systems.
- A company-wide educational curriculum on the ethical development of cognitive technologies.
- The creation of the IBM Cognitive Ethics and Society research program, a multi-disciplinary research program for the ongoing exploration of responsible development of AI systems aligned with our personal and professional values.
- Participation in cross-industry, government and scientific initiatives and events around AI and ethics, such as the White House Office of Science and Technology Policy AI workshops, the International Joint Conference on Artificial Intelligence, and the conference of the Association for the Advancement of Artificial Intelligence.
- Regular, ongoing IBM-hosted engagements with a robust ecosystem of academics, researchers, policymakers, NGOs and business leaders on the ethical implications of AI.

Conclusion

“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.

IBM Global Services
Route 100
Somers, NY 10589
USA.

Produced in the United States of America
September 2016