

BUILDING TRUST IN ARTIFICIAL INTELLIGENCE Author(s): Francesca Rossi

Source: *Journal of International Affairs*, Vol. 72, No. 1, THE FOURTH INDUSTRIAL REVOLUTION (Fall 2018/Winter 2019), pp. 127-134

Published by: Journal of International Affairs Editorial Board

Stable URL: <https://www.jstor.org/stable/10.2307/26588348>

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at <https://about.jstor.org/terms>



Journal of International Affairs Editorial Board is collaborating with JSTOR to digitize, preserve and extend access to *Journal of International Affairs*

JSTOR

BUILDING TRUST IN ARTIFICIAL INTELLIGENCE

Francesca Rossi

What is Artificial Intelligence?

Artificial Intelligence (AI) is a scientific discipline aimed at building machines that can perform many tasks that require human intelligence. AI started more than 60 years ago, and includes two main areas of research. One is based on rules, logic, and symbols; it is explainable; and it always finds a correct solution for a given problem, if that problem has been correctly specified. However, it can be used only when all possible scenarios for the problem at hand can be foreseen.

The other area of research is based on examples, data analysis, and correlation. It can be applied in cases where there is an incomplete or ill-defined notion of the problem to be solved. However, this type of AI requires a lot of data, is usually less explainable, and there is always a small margin of error. These two lines of research and ways of thinking about AI are increasingly being combined in order to maximize the advantages of both and to mitigate their drawbacks.

In recent years, many successful applications of AI have been built, mainly because of the convergence of improved algorithms, vast computing power, and massive amounts of data. This provides AI systems with human-level perception capabilities, such as speech-to-text, text understanding, image interpretation, and others, for which machine-learning methods are suitable. These abilities make it possible to deploy AI systems in real-life scenarios that typically have a high degree of uncertainty. Still, current consumer-oriented AI applications where a service is provided to users—from navigation systems to voice-activated “smart” homes—barely scratch the surface of the tremendous opportunity that AI represents for businesses and other institutions.

The main purpose of what can be called enterprise AI is to augment humans’ capabilities and to allow humans to make better—that is, more informed and grounded—decisions. At this point, AI and humans have very complementary

capabilities, and it is when their capabilities are combined that we find the best results. Typical applications in enterprise AI are decision-support systems for doctors, educators, financial service operators, and a host of other professionals who need to make complex decisions based on lots of data.

A Problem of Trust

It is easy to see that AI will become pervasive in our everyday life. This will certainly bring many benefits in terms of scientific progress, human wellbeing, economic value, and the possibility of exploring solutions to major social and environmental problems. However, such a powerful technology also raises some concerns, such as its ability to make important decisions in a way that humans would perceive as fair, to be aware and aligned to human values that are relevant to the problems being tackled, and the capability to explain its reasoning and decision-making. Since many successful AI techniques rely on huge amounts of data, it is important to know how data are handled by AI systems and by those who produce them.

These concerns are among the obstacles that hold AI back or that cause worry for current AI users, adopters, and policymakers. Issues include the black-box nature of some AI approaches, the possible discriminatory decisions that AI algorithms may make, and the accountability and responsibility when an AI system is involved in an undesirable outcome.

Without answers to these questions, many will not trust AI, and therefore will neither fully adopt it nor make use of its positive capabilities. According to a new study by IBM's Institute for Business Value, 82 percent of all enterprises and 93 percent of high-performing enterprises are now considering or moving ahead with AI adoption, attracted by the technology's ability to drive revenues, improve customer service, lower costs, and manage risk.¹ However, although they realize the huge benefits of this technology, 60 percent of those companies fear liability issues and 63 percent say they lack the skills to harness AI's potential, according to the same study.²

High-Level Principles for AI

Both researchers and policy thinkers are wrestling with the previously referenced questions. As a result of this ongoing discussion, several high-minded guiding principles about AI design, development, and usage have been defined and publicly evaluated:

- *IBM's* principles of trust and transparency:^{3,4} AI should augment human intelligence rather than replace it, trust is key to adoption, and data policies should be transparent.

- **Google's** principles on AI:⁵ AI should protect privacy and be socially beneficial, fair, safe, and accountable to people.
- The **Asilomar's** AI principles:⁶ Drafted at the 2017 Asilomar Conference, these 23 principles cover research, ethics, and values in AI, in addition to long-term issues. The principles have been signed by 1,273 researchers and 2,541 other interested parties, including Elon Musk and the late Stephen Hawking.
- The tenets of the **Partnership on AI (PAI)**:⁷ Eight tenets for an open and collaborative environment to discuss AI best practices, the social responsibility of companies delivering AI, AI explainability, and trust. Every partner that wants to join the PAI needs to sign onto these tenets.
- The **AI4PEOPLE** principles and recommendations:⁸ Concrete recommendations for European policymakers to facilitate the advance of AI in Europe.
- The **World Economic Forum's** principles for ethical AI:⁹ Five principles that cover the purpose of AI, its fairness and intelligibility, data protection, the right for all to exploit AI for their wellbeing, as well as the opposition to autonomous weapons.
- The **Institute of Electrical and Electronics Engineers** general principles:¹⁰ a set of principles that place AI within a human rights framework with references to wellbeing, accountability, corporate responsibility, value by design, and ethical AI.

Practical Implementation of High-Level Principles

Such principles are an important first step, but these conversations should be followed by concrete action to implement viable solutions.

Explainability

Companies and users want AI systems that are transparent, explainable, ethical, properly trained with appropriate data, and free of bias. Yet too often, commercially available AI systems are an opaque black box, offering users scarce visibility about the underlying data, processes, and logic that lead to the system's decisions. The most successful machine-learning approaches, such as those based on deep learning, are non-transparent and do not provide easy access into their decision-making. This makes explainability an outstanding challenge, although some attempts to demystify the technology are underway, including OpenScale from IBM.¹¹

Bias Awareness and Mitigation

Bias detection and mitigation are also fundamental in achieving trust in AI. Bias can be introduced through training data, when it is not balanced and inclusive

enough, but it can also be injected in the AI model in many other ways. Moreover, among the many notions of fairness, it is important to choose the most appropriate given the specific application context. It is also important to help developers become aware of what is available and can be used in current AI systems because of the abundance of bias metrics, notions of fairness, and bias mitigation and detection algorithms. The global community of data scientists and developers can and should continue to improve upon these capabilities in a collaborative way. To that end, IBM has made available to the open-source community a toolkit called “AI Fairness 360” to help developers and data scientists check for and mitigate bias in AI models using bias-handling solutions, and supporting them with guidelines, datasets, tutorials, metrics, and algorithms.¹²

Trusting AI Producers

Trust in the technology should be complemented by trust in those producing the technology. Yet such trust can only be gained if companies are transparent about their data usage policies and the design choices made while designing and developing new products. If data are needed to help AI make better decisions, it is important that the human providing the data is aware of how his/her data are handled, where they are stored, and how they are used. Regulations such as the General Data Protection Regulation (GDPR) in Europe provide some fundamental rights over personal data.¹³ Besides performance and accuracy, bias definition and detection and mitigation methods should also be communicated clearly, and explainability capabilities described and made accessible to all users.

The good news is that the industry is beginning to offer such solutions. For instance, research scientists and engineers at IBM have released a set of trust and transparency capabilities for AI, designed around three core principles: explainability, fairness, and traceability. These software tools provide explainability and bias detection on AI models in real time, detecting potentially unfair outcomes and automatically recommending data to add to the model to help mitigate bias.

Imagine an insurance company searching for unintended bias in their AI-powered claim fraud detection process. Using these tools, the company could flag discrepancies between normal and actual approval rates, identify any bias affecting the decision, and highlight factors that may influence why a claim was denied. The toolkit also shows a measure of the confidence that the system has in a recommendation and the factors behind that confidence level. The system would also automatically recommend adding certain kinds of data to help reduce instances of bias moving forward.

Additionally, businesses operating in heavily regulated industries often require extensive information about the decision processes of their AI systems.

The ability to track the accuracy, performance, and fairness of their applications, and of recording this information, can provide that level of detail for compliance, accountability, or customer-service purposes. To this end, IBM has proposed the idea of an “AI factsheet”, where developers should record all design decisions and performance properties of the developed AI system, from the bias handling algorithms, to the training datasets, to the explainability tools, etc.¹⁴ Also, to help developers and designers think about these issues, IBM has released a booklet, called “Everyday Ethics for Artificial Intelligence”, to raise the awareness of developers and designers on these topics and help them to think and find solutions to trust-related capabilities in their everyday job.¹⁵

Driving and Facilitating Trusted AI


Strong collaboration with policymakers and regulators is also needed. In the EU, the European Commission is taking a multi-pronged approach to fostering the responsible development and deployment of AI. In addition to public investment in research and the promotion of public-private partnerships, the EU has brought together experts from various disciplines in a European AI High Level Expert Group—of which I am a member—tasked with developing ethical guidelines that will consider principles such as values, explainability, transparency, and bias, as well as with recommending policies that include funding and infrastructures to support AI in Europe. In the United States, some have called for rules to constrain the design and use of AI, although a comprehensive approach is not yet in place to understand what the right approach might be for the American environment.

Besides hard laws, there are many ways a powerful technology such as AI can be directed toward beneficial impacts, such as standards, the proliferation of best practices and guidelines, and incentives. I believe that all of these tools should be used to achieve a proactive and participatory approach where both the AI producers and AI users can have a voice.

An example of concrete recommendations to policymakers is included in the previously-mentioned AI4PEOPLE white paper that I co-authored, where high-level principles and values are mapped into concrete recommendations for potential actions that EU policymakers can follow.¹⁶ These recommendations cover issues as wide-ranging as: assessing the capabilities of the technology and the institutions using it; the need for explainability and transparency, possibly supported by auditing mechanisms; formulating redress or compensation processes; the need for appropriate metrics for AI trustworthiness; developing a new EU oversight agency responsible for the protection of public welfare through the evaluation and supervision of AI; creating financial incentives for cross-disciplinary activities and ethical AI; and supporting education curricula that cover the social and legal impacts of

AI.

Conclusion

AI is a powerful technology that will have immense positive impacts on our lives. However, to fully gauge its potential benefits, we need to build a system of trust, both in the technology and in those who produce it. Issues of bias, explainability, data handling, transparency on data policies, and design choices should be addressed in a responsible and open way. By infusing AI products with robust and proven bias detection and mitigation capabilities, as well as the ability to explain how their decisions are made, AI developers can bridge the trust gap and create an effective path for economic growth and societal benefit. With the development of several high-level principles to guide AI toward a positive impact, it is time to put such principles to work and create robust implementation mechanisms. Only a holistic, multi-disciplinary, and multi-stakeholder approach can build such a system of trust, which must include AI makers, AI users, and policymakers. This system could ensure issues are identified, discussed, and resolved in a cooperative environment. It is this kind of interdisciplinary and cooperative approach that will produce the best solutions and is most likely to lead to a comprehensive environment for trustworthy AI. 

Francesca Rossi is the AI Ethics Global Leader at IBM Research, Professor of Computer Science at the University of Padova, Italy (on leave), and Member of the European Commission's High-Level Expert Group on Artificial Intelligence.

NOTES

¹ IBM Institute for Business Value, "Shifting toward Enterprise-grade AI: Resolving data and skills gaps to realize value" (IBM, 2018), <https://www-935.ibm.com/services/us/gbs/thoughtleadership/enterpriseai/>.

² Ibid.

³ "Trusted AI For Business" (IBM, 2018), <https://www.ibm.com/watson/ai-ethics/>.

⁴ "Trusted AI" (IBM, 2018), <https://www.research.ibm.com/artificial-intelligence/trusted-ai/>.

⁵ Sundar Pichai, "AI at Google: our principles" (Google, 7 June 2018), <https://www.blog.google/technology/ai/ai-principles/>.

⁶ "Asilomar AI Principles" (Future of Life Institute, 2017), <https://futureoflife.org/ai-principles/>.

⁷ "Tenets" (Partnership on AI, 2018), <https://www.partnershiponai.org/tenets/>.

⁸ "AI4People's Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations" (AI4PEOPLE, November 2018), <http://www.eismd.eu/wp-content/uploads/2018/11/Ethical-Framework-for-a-good-AI-Society.pdf>.

⁹ Rob Smith, "5 core principles to keep AI ethical" (World Economic Forum, 19 April 2018), <https://www.weforum.org/agenda/2018/04/keep-calm-and-make-ai-ethical/>.

¹⁰ "IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous

Systems” (Institute of Electrical and Electronics Engineers, 2018), https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_general_principles.pdf.

¹¹ “AI OpenScale” (IBM, 2018), <https://www.ibm.com/cloud/ai-openscale>.

¹² Kush Varshney, “Introducing AI Fairness” (IBM, 19 September 2018), <https://www.ibm.com/blogs/research/2018/09/ai-fairness-360/>.

¹³ EUGDPR Information Portal (2018), <https://eugdpr.org/>.

¹⁴ Aleksandra Mojsilovic, “Factsheets for AI Services” (IBM, 2018), <https://www.ibm.com/blogs/research/2018/08/factsheets-ai/>.

¹⁵ Adam Cutler, Milena Pribic, and Lawrence Humphrey, “Everyday Ethics for Artificial Intelligence” (IBM, 2018), <https://www.ibm.com/watson/assets/duo/pdf/everydayethics.pdf>.

¹⁶ AI4PEOPLE (2018).

