ACCOUNTABILITY IN AI

Promoting Greater Social Trust

Theme Paper for the G7 Multi-stakeholder Conference on Artificial Intelligence: Enabling the Responsible Adoption of AI
December 6, 2018

Authors:
Dr. Jason Millar, University of Ottawa
Brent Barron, CIFAR
Dr. Koichi Hori, University of Tokyo
Rebecca Finlay, CIFAR
Kentaro Kotsuki, IICP
Dr. Ian Kerr, University of Ottawa, Chair CIFAR AI & Society Council

CIFAR

情報通信政策研究所
Institute for Information and Communications Policy

Acknowledgements and Authors Note

We gratefully acknowledge the support of CIFAR and the Ministry of Internal Affairs and Communications of the Government of Japan as well as the many stakeholders in Canada and Japan who provided feedback and advice on this paper. Please note that the ideas expressed herein are either those of the authors or were provided through the stakeholder consultation. They are not those of the governments of Canada or Japan.

Executive Summary

This paper was developed at the request of the Government of Canada to support the G7 Multi-stakeholder Conference on Artificial Intelligence: Enabling the Responsible Adoption of AI on December 6, 2018. Co-leads from Canada and Japan developed this paper on accountability, the intent of which is to provide a starting point for discussions on the topic of Accountability in AI: Promoting Greater Social Trust at the conference. This paper and the discussion builds on work that started at the 2016 Takamatsu ICT Ministerial Meeting and led, most recently, to the Charlevoix Common Vision for the Future of Artificial Intelligence[1].

This paper is organized into two sections. The first provides information on work to date in this domain and sets out various concepts and distinctions worth noting when thinking about accountability and trust in AI. The second section reports on the consultation process and discusses potential actions for different stakeholder groups for the future.

Seven questions, organized under three broad headings, are proposed for framing the discussions at the conference:

Principles

1. What are some shared principles for Artificial Intelligence (AI) accountability in all sectors?
2. How do we determine which AI systems require more rigorous accountability regimes for their appropriate governance?

Development

3. Given that trust can be misplaced—individuals can over- and under-trust AI—how can accountability regimes promote the development of *trustworthy AI that is appropriately trusted*?
4. How do we balance accountability with innovation so that the benefits of AI are responsibly and inclusively secured?

Instruments

5. How can we ensure a representative and diverse plurality of voices and perspectives in the development of international and national accountability regimes for AI?
6. What mechanisms (regulatory vs. non-regulatory) are most appropriate to govern various applications of algorithmic decision-making?
7. What role should different stakeholders (e.g. governments; international organizations; private developers, service providers and users; the legal system; etc.) play in ensuring accountability in AI, and coordination across jurisdictional and cultural boundaries?

[1] https://g7.gc.ca/wp-content/uploads/2018/06/FutureArtificialIntelligence.pdf

Work to Date

Introduction

With the development and proliferation of AI systems, there is an urgent need to address questions of accountability. However, there is a "lack of consensus among the broader community regarding what a 'solutions toolkit' would look like."[2] This paper surveys the topic of accountability in AI and its link to trust, proposes some key definitions and distinctions, and provides some considerations for future discussions and potential actions among G7 members, other countries, and stakeholders worldwide.

The term artificial intelligence (AI) encompasses a broad range of technologies and approaches. Two general approaches to AI are worth distinguishing. One approach uses predefined models to accomplish goals; the other relies on machine learning to train a system to accomplish goals. There are two well-known techniques in machine learning. To define them at a very high level, they are deep learning, which uses very large artificial neural networks, and reinforcement learning, which uses a reward and punishment structure. The intent of this paper is to discuss accountability as it applies broadly to AI, while recognizing that certain ethical issues that have become associated with AI, most notably explainability, relate most directly to deep learning.

AI research has advanced rapidly in the past decade. Success in the lab has led to the proliferation of AI-based systems in certain sectors of society. Because of its ability to operate on massive data sets with speed, precision and accuracy that outpace human capacities, AI is beginning to be applied, or is being contemplated, in healthcare, transportation, law and order, defense, finance—virtually every sector of the economy—to support and in some cases substitute human analysis and decision-making. These capabilities position AI to deliver great benefits to society.

As with any new technology, we are learning that deploying AI beyond the lab might create risks for individuals and societies, raising concerns about accountability. A few examples that help to illustrate follow. AI that is trained on biased data sets can entrench and proliferate those biases in its outputs, leading to discriminatory applications[3]. In practice, many deep learning systems function largely as "black-boxes," and so their behaviour can be difficult to interpret and explain, raising concerns

Key Term: Artificial Intelligence (AI)

"[AI is] about making computers that can help us that can do the things that humans can do but our current computers can't" – Yoshua Bengio

"The field of computer science dedicated to solving cognitive problems commonly associated with human intelligence, such as learning, problem solving, and pattern recognition." – Amazon

"It is the science and engineering of making intelligent machines, especially intelligent computer programs. It is related to the similar task of using computers to understand human intelligence, but AI does not have to confine itself to methods that are biologically observable." – John McCarthy

[2] WWW Foundation. (2017). "Algorithmic Accountability: Applying the Concept to Different Country Contexts." World Wide Web Foundation. Online: https://webfoundation.org/docs/2017/07/Algorithms_Report_WF.pdf, 5.

[3] Sharkey, N. (2018). "The Impact of Gender and Race Bias in AI." ICRC: Humanitarian Law & Policy. Online: http://blogs.icrc.org/law-and-policy/2018/08/28/impact-gender-race-bias-ai/.

over explainability, transparency, and human control[4]. Moreover, AI systems may have multiple components (code, sensors, data assets, etc.), any of which may malfunction, further complicating how accountability is determined. Finally, owing to the way humans often perceive AI as "superior" in its abilities, they can over-trust it[5]. These examples are not exhaustive. As we learn more about AI and its unique characteristics, the list of potential harms is evolving. An understanding of these potential harms is beginning to be incorporated into governmental thinking on AI[6]. Indeed, systematic research into the ethical implications of AI is progressing steadily both inside and outside of academia. Some of that emerging research focuses specifically on helping policy makers and engineers anticipate and address ethical issues related to AI, including accountability[7].

Anticipating and addressing these potential risks is urgent. These systems are often opaque and complex, and their potential impact is broad. Coupled with their potential use in critical, high-stakes decision contexts (e.g. judicial reasoning, healthcare, warfare, financial transactions), their potential impact is significant. For example, a routine software update to a traffic routing algorithm controlling an automated and connected mobility system could quickly redistribute risks among millions of people within the system. Determining who ought to face greater risks within a mobility system is a weighty task with broad implications[8]. The process by which we ought to make that decision, as well as the responsibility for that decision and its systemic consequences, may exceed the capabilities of existing regimes (torts, consumer protection, etc.).[9]

Though there is clearly potential to do harm by deploying AI in some contexts, we should be measured in our concern. In many cases, the negative societal impacts of status quo (i.e. non-algorithmic) systems are not interrogated as intensely as AI systems[10]. In other words, it is important to understand the risks posed by AI as well as the risks posed by the status quo.

As is commonly the case, the pace of technical innovation is outpacing our policy responses with respect to accountability. Failing to establish clear guidance related to accountability could undermine trust among both experts and the public, potentially limiting the benefits of AI. At the same time, it is important to note that the goal cannot simply be to increase levels of trust in AI. This is because we can over- or under-trust an automated system. We under-trust when inaccurate assumptions (i.e. fears, misinformation) about AI prevent us from trusting it, potentially depriving us of the benefits it might produce. On the flipside, we over-trust a system when, for example, we mistakenly believe (and trust) that a system is capable of performing certain tasks that it is not. The unfortunate accidents caused by autonomous vehicles can be seen as cases of over-trust: in each case the human driver falsely believed that the automated system in control of the driving was capable of performing at a level at which it was

[4]Doshi-Velez, F., Kortz, M. (2017). "Accountability of AI Under the Law: The Role of Explanation." Berkman Klein Center Working Group on Explanation and the Law, Berkman Klein Center for Internet & Society working paper. Online: http://nrs.harvard.edu/urn-3:HUL.InstRepos:34372584

[5]Levin, S. (2018). "Tesla Fatal Crash: "Autopilot" Mode Sped Up Car Before Drive Killed, Report Finds." The Guardian. (8 June). Online: https://www.theguardian.com/technology/2018/jun/07/tesla-fatal-crash-silicon-valley-autopilot-mode-report; and Angwin, J. et al. (2016). "Machine Bias." ProPublica. Online: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[6]Cath et al. (2017). "Artificial Intelligence and the 'Good Society': The US, EU, and UK Approach." Science and Engineering Ethics 24(2): 505–528.

[7]Reisman, D. et al. (2018). "Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability." AI Now Institute. Online: https://ainowinstitute.org/aiareport2018.pdf; WWW Foundation; Doshi-Velez.; Jones, M., Millar, J. (2017). "Hacking Metaphors in the Anticipatory Governance of Emerging Technology: The Case of Regulating Robots." In R. Brownsword, E. Scotford & K. Yeung (Eds.) The Oxford Handbook on the Law and Regulation of Technology. (Oxford: Oxford University Press).

[8]Millar, J. (2017). "Ethics Setting for Autonomous Vehicles." In P. Lin, R. Jenkins, K. Abney & G.A. Bekey (Eds.) Robot Ethics 2.0. (Oxford: Oxford University Press).

[9]Millar, J., Kerr, I. (2016). "Delegation, Relinquishment, Responsibility: The Prospect of Expert Robots." In Ryan Calo, Michael Froomkin & Ian Kerr. (Eds.) Robot Law. (Cheltenham, UK: Edward Elgar Press).

[10] Cowgill, B. (2018). "The Impact of Algorithms on Judicial Discretion: Evidence from Regression Discontinuities." Working Paper. Online: http://www.columbia.edu/~bc2656/workingpapers.html

not capable of. Thus, our aim could be to encourage appropriate levels of trust in AI, with accountability regimes taking the nuances of over and under trust into account[11].

Finally, with the progress of AI networking where AI systems are connected to other systems, over the Internet or other information and communication networks, it will become more difficult to identify both the causes of issues as well as where the responsibility for them lies. In order to foster trust in AI, it will be important to build on a set of shared principles that clarify the roles and responsibilities for each stakeholder in the network including developers, service providers and end users in the research, development and use of AI.

Accountability, Trust and Trustworthiness

Broadly speaking, accountability is the foundation of trust in society. Accountability is about a clear acknowledgement and assumption of responsibility and "answerability" for actions, decisions, products and policies. Currently, three "senses" of accountability related to AI exist in the literature, each pointing to a different locus for action. In the first sense, accountability is a feature of the AI system itself[12]. Building explainability into the AI systems would partially address the AI's accountability in this sense. The second sense of accountability focuses on determining which individuals or groups are accountable for the impact of algorithms or AI[13]. In this sense, accountability is somewhat narrowly associated with determining who is most responsible for what effect within the sociotechnical system. Finally, and perhaps most broadly, accountability is seen as a feature of the broader sociotechnical system that develops, procures, deploys and uses AI[14]. For example, AI Now proposes an Algorithmic Impact Assessment framework (similar to a Privacy Impact Assessment) as a means of building accountability into the broader sociotechnical system in which AI is deployed, only part of which would include responsibility determinations. [15]Along similar lines, the World Wide Web (WWW) Foundation identifies principles of algorithmic accountability, including: fairness, explainability, auditability, responsibility, and accuracy.[16]

All three senses of accountability are being actively researched and developed.

The WWW Foundation describes a "critical" distinction between "algorithmic accountability—the responsibility of algorithm designers to provide evidence of potential or realised harms," and "algorithmic justice—the ability to provide redress from harms."[17] Their reason for making this distinction is the worry that focusing on redress as a means of addressing accountability distracts from a critical opportunity available to algorithm designers and engineers to anticipate harms before the AI is deployed. While taking this advice to heart, one must also be careful not to place too much emphasis on the responsibility of algorithm designers to anticipate harms, which could distract from a broader approach for addressing accountability in AI.

The above trust and accountability considerations point to a useful distinction between trusting a

[11] Levin, S. (2018).
[12] Doshi-Velez & Kortz; Villani, C. (2018). "For a Meaningful Artificial Intelligence: Towards A French and European Strategy." Ai for Humanity. Online: https://www.aiforhumanity.fr/pdfs/MissionVillani_Report_ENG-VF.pdf
[13] House of Lords, Select Committee on Artificial Intelligence. (2018). "AI in the UK: ready, willing and able?", Yeung, K. quoted at 96. Online: https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf; Larus, J. et al. (2018). "When Computers Decide: European Recommendations on Machine-Learned Automated Decision Making." Informatics Europe & EUACM. Online: http://www.informatics-europe.org/component/phocadownload/category/10-reports.html?download=74:automated-decision-making-report
[14]Reisman, D. et al. (2018); House of Lords, 35.
[15] Reisman, D. et al. (2018).
[16] WWW Foundation.
[17]Ibid, 5.

system and the trustworthiness of a system[18]. Trusting a system appropriately means having a justified level of trust in a system, that is, having just the right amount of trust in it. Thus, the trustworthiness of a system can be defined as the extent to which a system can reliably perform or fulfill its designated purpose as expected. For example, news stories shared on social media platforms are frequently trusted at a level higher than they should be, because some news stories are inaccurate and therefore are not trustworthy. Similarly, scientific publications are often trusted less than their trustworthiness would justify, and thus are often under-trusted. As a final example, people who trust flying in airplanes are trusting appropriately, because by all measures, air travel is a very trustworthy mode of transportation. When it comes to AI, various factors can cause people to not trust otherwise trustworthy AI. By developing robust accountability regimes for AI systems, including the broader surrounding sociotechnical systems, appropriate trust in AI would be promoted among experts and the public.

Transparency is often mentioned in discussions of AI accountability, because transparency allows for greater scrutiny of an AI system. However, accountability does not necessarily increase or improve simply by increasing transparency. In the absence of robust processes, principles, and frameworks, transparency alone not sufficient to ensure greater accountability.

Another challenge for accountable AI is that AI is portable across borders. It is developed and deployed in multiple jurisdictions, and in ways that cross international and cultural boundaries. Distribution and movement of digital assets is difficult to constrain. This has the effect of complicating trust, for example, when AI that is developed with one set of cultural assumptions embedded into it, is deployed in a "foreign" cultural context, where trust-building norms differ. It also complicates individual jurisdictional responses, since an AI might or might not be built to respect the local laws and appropriate cultural norms. The difficulties of dealing with cross-jurisdictional issues are not new, characterizing a number of issues in the digital age, privacy being chief among them. As we have seen with the recent European General Data Protection Regulation (GDPR), cross-jurisdictional solutions require multi-stakeholder input and would benefit from multi-lateral coordination. This coordination could not only ensure that an AI is functioning within the legal constraints of multiple jurisdictions, but also that it is functioning safely and in a trustworthy manner.

Finally, more research is needed to better support decision-making related to:

- bias
- explainability in AI
- ethics in engineering/design processes[19]
- effective public and multi-stakeholder engagement strategies for accountability in AI
- effective policy options
- ethical issues in AI
- legal analyses
- effective AI monitoring and audit strategies and techniques across multiple technologies and systems
- computational journalism , etc.[20]

[18] Aitken, M., Cunningham-Burley, S., Pagliari, C. (2016). "Moving from Trust to Trustworthiness: Experiences of Public Engagement in the Scottish Health Informatics Programme." Science and Public Policy 43(5): 713-723; Also, see Kaminski, M. et al. (2017). "Averting Robot Eyes." Maryland Law Review 76(4):

[19] Future of Humanity Institute. (2018). "The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation." Online: https://maliciousaireport.com.

[20] See WWW Foundation.

International Activity

Various international policies, programs, centres, and activities have been launched to address the development of robust and global AI accountability. Some of the gaps in knowledge they are tackling include how to:

(i) secure public sector input on the design of appropriate accountability regimes for AI;
(ii) support the development and maintenance of robust and global AI accountability regimes;
(iii) develop principles or working definitions of different "senses" of accountability;
(iv) develop algorithmic literacy strategies for informing and educating various stakeholder groups on the nature and impacts of algorithms they are subjected to;
(v) build consensus among the broader community regarding what a "solutions toolkit" would look like[21];
(vi) develop key indicators for Algorithmic Accountability and Algorithmic Justice; and,
(vii) establish clarity and/or agreement on the roles of various actors and stakeholders in ensuring accountability in AI. [22]

The following are some examples of work underway in standards and principles development, as well as individual jurisdictional approaches.

Declarations of Principles

Governments and other multi-stakeholder groups at the national, regional and municipal levels are declaring principles that will guide various aspects of AI development, procurement and use.[23] Additionally, a number of private organizations have introduced principles-based frameworks for the responsible adoption of AI. These include Google, SAP, and Microsoft.[24]

Japan

The Conference of Advisory Experts of Japan's Ministry of Internal Affairs and Communications has drafted AI R&D Principles to promote the societal and economic benefits of AI while mitigating risks, such as transparency and loss of control. The Conference's overarching vision is that of a Wisdom Network Society:

> "...a society where, as a result of the progress of AI networking, humans live in harmony with AI networks, and data/information/knowledge are freely and safely created, distributed, and linked to form a wisdom network, encouraging collaborations beyond space among people, things, and events in various fields and consequently enabling creative and vibrant developments."[25]

The principles for realizing this vision include collaboration, transparency, controllability, safety, security, privacy, ethics, user assistance, and accountability.

[21] Ibid.
[22] Latonero, M. (2018). "Governing Artificial Intelligence." Data & Society. Online: https://datasociety.net/output/governing-artificial-intelligence/
[23] See, for example: "Draft AI R&D GUIDELINES for International Discussions." Online: http://www.soumu.go.jp/main_content/000507517.pdf
[24] Google AI. "Responsible AI Practices." https://ai.google/education/responsible-ai-practices; SAP. "SAP's Guiding Principles for Artificial Intelligence." https://news.sap.com/2018/09/sap-guiding-principles-for-artificial-intelligence/; Microsoft. "Microsoft AI Principles." https://www.microsoft.com/en-us/ai/our-approach-to-ai.
[25] Ibid.

Building on that work, the Conference has introduced Draft AI Utilization Principles[26], which puts forward principles through the three pillars of promoting benefits, mitigating harms, and building trust:

1. Principle of proper utilization
2. Principle of data quality
3. Principle of collaboration
4. Principle of safety
5. Principle of security
6. Principle of privacy
7. Principles of human dignity and individual autonomy
8. Principle of fairness
9. Principle of transparency
10. Principle of accountability

In May 2018, the Cabinet Office of Japan began discussions toward the formulation of the social principles for human-centric AI, which will be basic principles for better social implementation and sharing of AI. The AI Social Principles will be finalized in March 2019.

Canada

The Montreal Declaration on the Responsible Development of AI, the result of a multi-stakeholder engagement process spearheaded by the Université de Montréal, seeks to outline "a series of ethical guidelines for the development of AI."[27] The first draft identifies seven key values to keep in mind when developing AI: "well-being, autonomy, justice, privacy, knowledge, democracy and accountability."

Standards Development

Several organizations (professional and otherwise) are working towards developing standards for the ethical development and use of AI. [28]

Institute of Electrical and Electronic Engineers (IEEE)

In 2016, IEEE, the world's largest professional engineering organization, established the Global Initiative on Ethics of Autonomous and Intelligent Systems. In its second version, this guiding document describes the ongoing work of various standards working groups that have since been established to address a number of sub-domains, including:

- data privacy;
- transparency of autonomous systems;
- a model process for addressing ethical concerns during system design;
- standards for ethically driven nudging for robotic, intelligent and autonomous systems; and,
- well-being metrics for ethical AI and autonomous systems.

[26] "Draft AI Utilization Principles." Online: http://www.soumu.go.jp/main_content/000581310.pdf

[27] "The Montreal Declaration for Responsible AI." Online: https://www.montrealdeclaration-responsibleai.com

[28] See, for example: IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, "Ethically Aligned Design: A Vision for Prioritizing Human Well-Being with Autonomous and Intelligent Systems." https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v2.pdf; and Canada's CIO Strategy Council: https://ciostrategycouncil.com/standards/new-projects/

International Standards Organization (ISO)

ISO has recently created a new technical subcommittee in the area of AI (SC 42), which is working to develop foundational standards as well as addressing issues related to safety and trustworthiness. SC 42 has created study groups on computational approaches and characteristics, trustworthiness, and use cases and applications.

These initiatives offer promising starting points and have the potential to contribute positive and significant results within their individual mandates; much can be learned and transferred from them. However, more work is needed to develop a fully articulated, robust and global AI accountability regime.

Below are examples of formal strategies undertaken by individual jurisdictions that may serve as precedents for other regions:

Government of Canada Directive on Automated Decision-Making

The Government of Canada is working towards releasing the first version of its Directive on Automated Decision-Making, which, in its current draft[29], sets out several requirements for AI development and use. Those include rules for performing Algorithmic Impact Assessments[30]; transparency and explainability; quality assurance; ensuring human intervention; recourse and reporting.

The European General Data Protection Regulation (GDPR)

True to its name, the GDPR a regulatory initiative that sets out general data protection rules aimed at protecting individuals' privacy within the EU. In addition to outlining rules concerning individual consent to data use, Articles 13-15 in particular set out what has been referred to as a "right to explanation" when algorithmic decision-making occurs. That is, individuals have a right to request information explaining the algorithmic logic used to render a decision when a system uses their personal data. Some have argued that this poses a barrier to AI innovation, both in terms of direct costs associated with manual reviews of algorithmic decisions, but also in terms of limiting potential performance of AI[31], whereas others see the GDPR as a move towards improving AI accountability.[32]

The NYC Automated Decision Systems Task Force

Billed as the first of its kind in the US, this nascent task force promises to "[recommend] a process for reviewing government automated decision systems, more commonly known as algorithms."[33] Their focus will be on ensuring that algorithms are "used appropriately and align with the goal of making New York City a fairer and more equitable place for all its residents."

[29] A Google Docs version of the draft Directive is publicly available for comment at: https://docs.google.com/document/d/1LdciG-UYeokx3U7ZzRng3u4T3IHrBXXk9JddjjueQok/edit#heading=h.hhc511vr644i

[30] This is also in the process of development.

[31] Wallace, N., Castro, D. (2018). "The Impact of the EU's New Data Protection Regulation on AI." Center for Data Innovation. Online: http://www2.datainnovation.org/2018-impact-gdpr-ai.pdf

[32] Doshi-Velez & Kortz.

[33] New York City Automated Decision Systems Task Force: https://www1.nyc.gov/site/adstaskforce/index.page

Looking Forward

Building on this brief overview of AI, accountability and activities underway, the following section intends to catalyse discussion at the December 6th conference and potential actions for the future. We begin with a short and non-exhaustive list of roles for potential stakeholder groups, as well as some suggested discussion topics and potential G7 leadership opportunities to be considered at the conference.

Roles for Multiple Stakeholders

Due to the complexity and intersectionality of issues related to AI and accountability, it will be critical that inclusive opportunities are created for diverse stakeholder groups to come together to move this work forward for the benefit of people worldwide. A number of different stakeholders could be engaged to provide role-specific input on the development and maintenance of robust global AI accountability regimes. Some examples are provided below.

Potential Roles:

Policymakers in National Governments

- Coordinate policy activity in national and international contexts.
- Promote responsible and inclusive government research in computer science, ethical robotics and AI engineering practices, and legal innovation.
- Promote the creation and verification of accountability, trustworthiness and other AI standards, both nationally and internationally, that account for the unique opacity of many AI systems, and the power of AI to have broad and rapid impacts on society.

Intergovernmental Organizations

- Provide a forum for convening international stakeholders to discuss high-level accountability strategies.
- Work towards policies and coordination mechanisms for addressing AI accountability and trust issues.

Policymakers in Sub-National Governments

- Convene relevant local stakeholders to develop responsible and inclusive solutions to AI-related challenges in sub-national jurisdictions.
- Provide opportunities for responsible and inclusive AI experimentation and share best practices with other jurisdictions.
- Support localized AI accountability and trustworthiness work that draws on region-specific law, economics, or culture.

Corporations and Other Data Owners

- Ensure appropriate levels of human control in the design and use of automated (algorithmic) decision-making.
- Implement transparent, meaningful ethics and accountability processes throughout the innovation lifecycle.
- Define and promote codes of conduct to support accountability.
- Engage regulators to help identify opportunities for responsible regulation, for example to help coordinate industry responses where externalities result from algorithmic decision-making.

Universities and Colleges

- Ensure AI Ethics is a core aspect of computer science and engineering curricula, and that coding literacy is a core aspect of social science and humanities curricula.
- Coordinate interdisciplinary research, workshops and meetings to further promote topics in the ethical engineering of robotics and AI.

Advocacy Groups and Public Interest Organizations

- Provide mechanisms for inclusive citizen engagement.
- Bridge knowledge sharing and dialogue between government and private sectors.
- Promote fair and open data sets for model training.

Foundations

- Invest in responsible and inclusive AI research and innovation.
- Provide platforms for researchers to help craft AI toolkits and evaluation frameworks.

Professional Regulatory Bodies and Organizations

- Develop Codes of Conduct and accountability mechanisms for licensed members, that account for members' unique ability to have broad and rapid impacts on society through the development, procurement, deployment and use of AI systems.

What We Heard

An early draft of this paper was placed online for public consultation. We received feedback from a number of individuals in Canada and Japan. Much of their feedback has been incorporated into the paper, but we have also attempted to present a summary below. Please note that these have been condensed or reworded, and are intended to represent the views of those consulted, not necessarily the views of the authors.

Multi-stakeholder Engagement

- Ensure that policymakers engage with deep technical experts.
- Encourage genuine diversity in engagement, particularly with marginalized communities and civil society.
- Ensure that conflicts of interest are managed, particularly among stakeholders that will benefit from widespread AI adoption.
- Pivot discussion from accountability to ethics more broadly.
- Consider the various governance and accountability intersections at international, national, regional, and municipal levels.
- Ensure that the perspectives of low- and middle-income countries are incorporated into decisions and proposals.
- Ensure that participation is not limited to beneficiaries or advocates of new technology.

Opportunities for Stakeholder Action

- Promote greater diversity in the technology workforce.
- Support public education and public ethics research.
- Develop data standards.
- Develop sectoral working groups by application.
- Ensure that regulatory and standards development processes are open and not restricted by stakeholders' financial resources.
- Consider a research ethics board or clinical trials style body for certain AI applications.
- Ensure that government staff at all levels have a sufficient understanding of AI to provide both oversight and identify opportunities for service modernization.
- Consider where existing policy, such as privacy, security, trade secrets, and copyright, may create barriers to accountability.
- Promote accessibility to the public by issuing all communications in plain language.
- Incorporate members of the public into any advisory or governance bodies.
- Develop funding support mechanisms so that vulnerable groups and youth can fully participate in discussions on an equal basis with industrial representatives.
- Support the promotion of trust in AI through the public commitment of organizations to specific principles, and further, through verification by third party compliance audits.
- While the free market may eventually arrive at an equilibrium of accountable AI, the transition period has the possibility of non-trivial harms, making a case for government involvement.

Other Considerations

- Pay close attention to scenarios in which human control of AI systems can be lost.
- Examine potential future AI scenarios, not just current challenges and opportunities
- Transparency is necessary but not sufficient.
- It is critical that new technology not change the fundamental role of civil society.
- Make clear distinctions between AI effects that can cause bodily harms versus financial harms and treat each appropriately.
- Accountability is a primarily social, rather than technical, challenge.
- AI deployment exists within a social architecture of cultural, legal, economic, and political contexts.
- Ensure that autonomous systems are accountable to people that are affected by these systems.
- Increasing autonomy of systems, particularly physical autonomy of robots, increases a variety of risk factors.
- Codes of conduct will not be sufficient to shape behaviour and constrain abuse; legal responses will be required.
- Non-technical solutions, such as disclosure requirements for large commercial systems, may be more productive than emerging technical research.
- The AI industry will need to ensure responsible activity so that its social license is maintained, particularly in sensitive fields, such as healthcare.
- Soft law, such as development and utilization guidelines and principles, can eventually become de facto hard law as these standards are increasingly viewed as baselines of responsibility to avoid negligence.
- Multinational coordination of standards would reduce obstacles to the responsible deployment of AI.
- Development and utilization of AI is in the early stage, and in order not to hamper innovation, the principles governing AI should be non-regulatory.
- Compensation may contribute to promoting accountability.
- Risk is estimated by multiplying the probability to the gravity of assumed loss. As risks are sometimes assessed only by the gravity of loss, it is necessary to properly assess the risk with consideration of the probability.
- There is a possibility that the assessed risks may differ depending on the cultures, and it may be necessary to set a different level of accountability for each culture.
- Preparing a mechanism for stopping the use of AI immediately when damage is caused by the use has the effect of preventing the spread of damage and leads to improvement of the trust.
- Consider promoting the trust including the foundation of an insurance system.
- Consider establishing an exemption clause like the aircraft accident investigation committee.
- Clarify what should be explained, to what extent explanation is required and what kind of explanation method is acceptable.

The G7 Looking Forward

Building on the discussion initiated by the 2016 G7 ICT Ministerial Meeting in Takamatsu, G7 members have undertaken studies on the potential social, economic, ethical, and legal issues raised by AI, as well as AI's socio-economic impact.

The G7 also recognises the need for further information sharing and discussion to deepen the understanding of the multi-faceted opportunities and challenges brought by AI. There are a number of potential roles that the G7 and other multi-lateral groups could play in the promotion of greater accountability in the AI sector. Some examples are provided below:

- Explore the potential for the emerging Canada-France international study group on AI to include other G7 members.
- Formally endorse an existing, or create a new declaration of, principles on ethical AI.
- Form a G7 working group to meet regularly and share best practices for different topics, including accountability frameworks and ethical AI use in government.
- Commit to a regular Multi-stakeholder Summit, such as December 6th, to examine emerging AI and accountability issues, in an open and inclusive forum.
- G7 commitment of support for national-level initiatives regarding accountability.

Discussion Questions for the G7 Meeting

The G7 Multi-stakeholder Conference on December 6th offers an opportunity for diverse stakeholder groups to come together to discuss some of these issues and to explore opportunities to move forward on the creation of robust and globally accountable AI. To begin to frame this discussion, we propose seven questions for discussion in the following themes:

Principles

1. What are some shared principles for Artificial Intelligence (AI) accountability in all sectors?
2. How do we determine which AI systems require more rigorous accountability regimes for their appropriate governance?

Development

3. Given that trust can be misplaced—individuals can over- and under-trust AI—how can accountability regimes promote the development of trustworthy AI that is appropriately trusted?
4. How do we balance accountability with innovation so that the benefits of AI are responsibly and inclusively secured?

Instruments

5. How can we ensure a representative and diverse plurality of voices and perspectives in the development of international and national accountability regimes for AI?
6. What mechanisms (regulatory vs. non-regulatory) are most appropriate to govern various applications of algorithmic decision-making?
7. What role should different stakeholders (e.g. governments; international organizations; private developers, service providers and users; the legal system; etc.) play in ensuring accountability in AI, and coordination across jurisdictional and cultural boundaries?

Conclusion

We are appreciative of the opportunity to provide this overview of AI and accountability in an effort to stimulate robust discussion at the December 6th conference. As the development of AI applications expands and accelerates, it is urgent and important for stakeholders to come together from all sectors, and across borders, to better understand what accountability means in an AI-enabled world and the implications for societal trust. We hope this survey of AI accountability is a useful resource for conference participants and others, stimulating future discussions and potential actions among G7 members, other countries and stakeholders worldwide.