Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI

PRINCIPLED ARTIFICIAL INTELLIGENCE:

Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI

Jessica Fjeld, Nele Achten, Hannah Hilligoss,
Adam Christopher Nagy, Madhulika Srikumar

BERKMAN
KLEIN CENTER
FOR INTERNET & SOCIETY
AT HARVARD UNIVERSITY

Table of Contents

2 Acknowledgements

1. **3 Introduction**
4 Executive Summary
7 How to Use these Materials
8 Data Visualization

2. **11 Definitions and Methodology**
11 Definition of Artificial Intelligence
12 Definition of Relevant Documents
14 Document Search Methodology
15 Principle and Theme Selection Methodology
18 Timeline Visualization

3. **20 Themes among AI Principles**
3.1 21 Privacy
3.2 28 Accountability
3.3 37 Safety and Security
3.4 41 Transparency and Explainability
3.5 47 Fairness and Non-discrimination
3.6 53 Human Control of Technology
3.7 56 Professional Responsibility
3.8 60 Promotion of Human Values

4. **64 International Human Rights**

5. **66 Conclusion**

6. **68 Bibliography**

Acknowledgements

This report, *Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI*, is part of the Berkman Klein Center for Internet & Society research publication series. The report and its authors have significantly benefitted from more than two years of research, capacity building, and advocacy executed by the Berkman Klein Center as co-leader of the Ethics and Governance of Artificial Intelligence Initiative in partnership with the MIT Media Lab. Through the Initiative, the Center has worked on pressing issues raised by the development and deployment of AI-based systems – including groundbreaking research on automation's impact on the online media landscape; fostering the growth of inclusive AI research and expertise in academic centers across the globe; and advocating for increased transparency surrounding the use of algorithms in criminal justice decision-making. Notably, the Center engages both private-sector firms and public policymakers in troubleshooting related legal, social, and technical challenges, and has consulted on the development of AI governance frameworks with national governments and intergovernmental bodies, such as the UN High-Level Committee on Programmes and the OECD's Expert Group on AI. *Principled Artificial Intelligence* has been influenced by these collaborative, multistakeholder efforts, and hopes to contribute to their continued success.

The report's lead author, Jessica Fjeld, is the assistant director of the Harvard Law School Cyberlaw Clinic, based at the Center; Nele Achten is an affiliate; Hannah Hilligoss a former Senior Project Coordinator and current HLS JD candidate; Adam Nagy a current Project Coordinator; and Madhulika Srikumar a former Cyberlaw Clinic student and current HLS LLM candidate. Thank you to Amar Ashar, Christopher Bavitz, Ryan Budish, Rob Faris, Urs Gasser, Vivek Krishnamurthy, Momin Malik, Jonathan Zittrain, and others in the Berkman Klein community for their generous input, support, and questions.

Thank you to Maia Levy Daniel, Joshua Feldman, Justina He, and Sally Kagay for indispensable research assistance, and to Jessica Dheere, Fanny Hidvégi, Susan Hough, K.S. Park, and Eren Sozuer for their collegial engagement on these issues and this project, as well as to all the individuals and organizations who contributed comments on the draft data visualization we released in summer 2019. An updated and final data visualization accompanies this report: thank you to Melissa Axelrod and Arushi Singh for their thoughtfulness and significant skill in its production.

1. Introduction

Alongside the rapid development of artificial intelligence (AI) technology, we have witnessed a proliferation of “principles” documents aimed at providing normative guidance regarding AI-based systems. Our desire for a way to compare these documents – and the individual principles they contain – side by side, to assess them and identify trends, and to uncover the hidden momentum in a fractured, global conversation around the future of AI, resulted in this white paper and the associated data visualization.

It is our hope that the Principled Artificial Intelligence project will be of use to policymakers, advocates, scholars, and others working on the frontlines to capture the benefits and reduce the harms of AI technology as it continues to be developed and deployed around the globe.

Executive Summary

In the past several years, seemingly every organization with a connection to technology policy has authored or endorsed a set of principles for AI. As guidelines for ethical, rights-respecting, and socially beneficial AI develop in tandem with – and as rapidly as – the underlying technology, there is an urgent need to understand them, individually and in context. To that end, we analyzed the contents of thirty-six prominent AI principles documents, and in the process, discovered thematic trends that suggest the earliest emergence of sectoral norms.

While each set of principles serves the same basic purpose, to present a vision for the governance of AI, the documents in our dataset are diverse. They vary in their intended audience, composition, scope, and depth. They come from Latin America, East and South Asia, the Middle East, North America, and Europe, and cultural differences doubtless impact their contents. Perhaps most saliently, though, they are authored by different actors: governments and intergovernmental organizations, companies, professional associations, advocacy groups, and multi-stakeholder initiatives. Civil society and multistakeholder documents may serve to set an advocacy agenda or establish a floor for ongoing discussions. National governments' principles are often presented as part of an overall national AI strategy. Many private sector principles appear intended to govern the authoring organization's internal development and use of AI technology, as well as to communicate its goals to other relevant stakeholders including customers and regulators. Given the range of variation across numerous axes, it's all the more surprising that our close study of AI principles documents revealed common themes.

The first substantial aspect of our findings are the **eight key themes** themselves:

- **Privacy.** Principles under this theme stand for the idea that AI systems should respect individuals' privacy, both in the use of data for the development of technological systems and by providing impacted people with agency over their data and decisions made with it. Privacy principles are present in 97% of documents in the dataset.
- **Accountability.** This theme includes principles concerning the importance of mechanisms to ensure that accountability for the impacts of AI systems is appropriately distributed, and that adequate remedies are provided. Accountability principles are present in 97% of documents in the dataset.
- **Safety and Security.** These principles express requirements that AI systems be safe, performing as intended, and also secure, resistant to being compromised by unauthorized parties. Safety and Security principles are present in 81% of documents in the dataset.
- **Transparency and Explainability.** Principles under this theme articulate requirements that AI systems be designed and implemented to allow for oversight, including through translation of their operations into intelligible outputs and the provision of information about where, when, and how they are being used. Transparency and Explainability principles are present in 94% of documents in the dataset.
- **Fairness and Non-discrimination.** With concerns about AI bias already impacting individuals globally, Fairness and Non-discrimination principles call for AI systems to be designed and used to maximize fairness and promote inclusivity. Fairness and Non-discrimination principles are present in 100% of documents in the dataset.
- **Human Control of Technology.** The principles under this theme require that important decisions remain subject to human review. Human Control of Technology principles are present in 69% of documents in the dataset.
- **Professional Responsibility.** These principles recognize the vital role that individuals involved in the development and deployment of AI systems play in the systems' impacts, and call on their professionalism and integrity in ensuring that the appropriate stakeholders are consulted and long-term effects are planned for. Professional Responsibility principles are present in 78% of documents in the dataset.
- **Promotion of Human Values.** Finally, Human Values principles state that the ends to which AI is devoted, and the means by which it is implemented, should correspond with our core values and generally promote humanity's well-being. Promotion of Human Values principles are present in 69% of documents in the dataset.

The second, and perhaps even more striking, side of our findings is **that more recent documents tend to cover all eight of these themes**, suggesting that the conversation around principled AI is beginning to converge, at least among the communities responsible for the development of these documents. Thus, these themes may represent the "normative core" of a principle-based approach to AI ethics and governance.[1]

However, we caution readers against inferring that, in any individual principles document, broader coverage of the key themes is necessarily better. Context matters. Principles should be understood in their cultural, linguistic, geographic, and organizational context, and some themes will be more relevant to a particular context and audience than others. Moreover, principles are a starting place for governance, not an end. On its own, a set of principles is unlikely to be more than gently persuasive. Its impact is likely to depend on how it is embedded in a larger governance ecosystem, including for instance relevant policies (e.g. AI national plans), laws, regulations, but also professional practices and everyday routines.

[1] Both aspects of our findings are observable in the data visualization (p. 8-9) that accompanies this paper.

One existing governance regime with significant potential relevance to the impacts of AI systems is international human rights law. Scholars, advocates, and professionals have increasingly been attentive to the connection between AI governance and human rights laws and norms,[2] and we observed the impacts of this attention among the principles documents we studied. 64% of our documents contained a reference to human rights, and five documents took international human rights as a framework for their overall effort. Existing mechanisms for the interpretation and protection of human rights may well provide useful input as principles documents are brought to bear on individuals cases and decisions, which will require precise adjudication of standards like "privacy" and "fairness," as well as solutions for complex situations in which separate principles within a single document are in tension with one another.

The thirty-six documents in the *Principled Artificial Intelligence* were curated for variety, with a focus on documents that have been especially visible or influential. As noted above, a range of sectors, geographies, and approaches are represented. Given our subjective sampling method and the fact that the field of ethical and rights-respecting AI is still very much emergent, we expect that perspectives will continue to evolve beyond those reflected here. We hope that this paper and the data visualization that accompanies it can be a resource to advance the conversation on ethical and rights-respecting AI.

[2] Hannah Hilligoss, Filippo A. Raso and Vivek Krishnamurthy, 'It's not enough for AI to be "ethical"; it must also be "rights respecting"', Berkman Klein Center Collection (October 2018) https://medium.com/berkman-klein-center/its-not-enough-for-ai-to-be-ethical-it-must-also-be-rights-respecting-b87f7e215b97.

How to Use these Materials

Data Visualization

The Principled AI visualization, designed by Arushi Singh and Melissa Axelrod, is arranged like a wheel. Each document is represented by a spoke of that wheel, and labeled with the sponsoring actors, date, and place of origin. The one exception is that the OECD and G20 documents are represented together on a single spoke, since the text of the principles in these two documents is identical.[3] The spokes are sorted first alphabetically by the actor type and then by date, from earliest to most recent.

Inside the wheel are nine rings, which represent the eight themes and the extent to which each document makes reference to human rights. In the theme rings, the dot at the intersection with each spoke indicates the percentage of principles falling under the theme that the document addresses: the larger the dot, the broader the coverage. Because each theme contains different numbers of principles (ranging from three to ten), it's instructive to compare circle size within a given theme, but not between then.

In the human rights ring, a diamond indicates that the document references human rights or related international instruments, and a star indicates that the document uses international human rights law as an overall framework.

[3] Note that while the OECD and G20 principles documents share a single spoke on the data visualization, for purposes of the quantitative analysis underlying this paper, they have been counted as separate documents.

PRINCIPLED ARTIFICIAL INTELLIGENCE

A Map of Ethical and Rights-Based Approaches to Principles for AI

Authors: Jessica Fjeld, Nele Achten, Hannah Hilligoss, Adam Nagy, Madhulika Srikumar

Designers: Arushi Singh (arushisingh.net) and Melissa Axelrod (melissaaxelrod.com)

HOW TO READ:

Date, Location
Document Title
Actor

COVERAGE OF THEMES:

Higher ← → Lower

● ● ● • • ○

Not referenced

◆ References International Human Rights

✷ Explicitly Adopts Human Rights Framework

The size of each dot represents the percentage of principles in that theme contained in the document. Since the number of principles per theme varies, it's informative to compare dot sizes within a theme but not between themes.

The principles within each theme are:

Privacy:
Privacy
Control over Use of Data
Consent
Privacy by Design
Recommendation for Data Protection Laws
Ability to Restrict Processing
Right to Rectification
Right to Erasure

Accountability:
Accountability
Recommendation for New Regulations
Impact Assessment
Evaluation and Auditing Requirement
Verifiability and Replicability
Liability and Legal Responsibility
Ability to Appeal
Environmental Responsibility
Creation of a Monitoring Body
Remedy for Automated Decision

Safety and Security:
Security
Safety and Reliability
Predictability
Security by Design

Transparency and Explainability:
Explainability
Transparency
Open Source Data and Algorithms
Notification when Interacting with an AI
Notification when AI Makes a Decision about an Individual
Regular Reporting Requirement
Right to Information
Open Procurement (for Government)

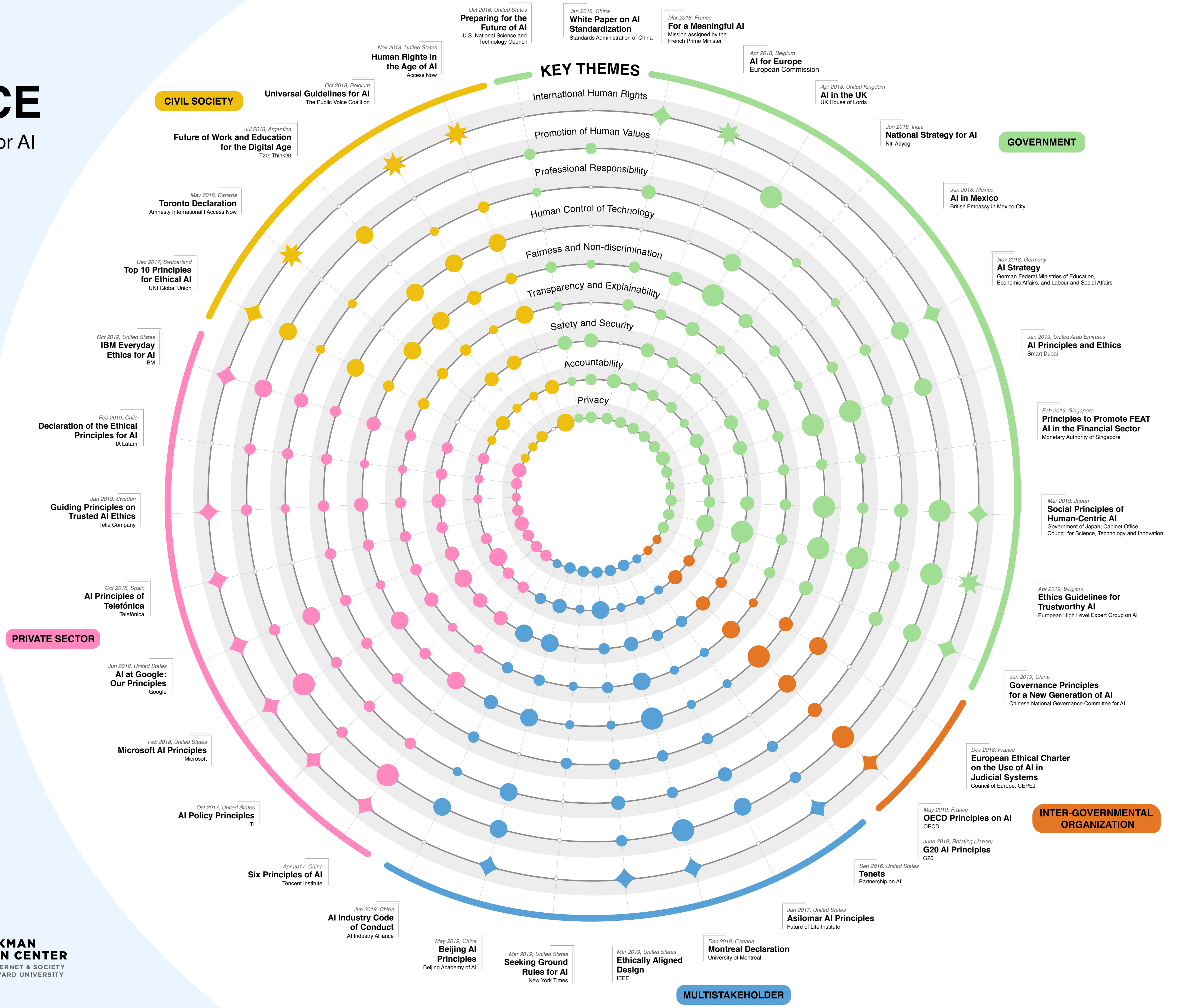
Fairness and Non-discrimination:
Non-discrimination and the Prevention of Bias
Fairness
Inclusiveness in Design
Inclusiveness in Impact
Representative and High Quality Data
Equality

Human Control of Technology:
Human Control of Technology
Human Review of Automated Decision
Ability to Opt out of Automated Decision

Professional Responsibility:
Multistakeholder Collaboration
Responsible Design
Consideration of Long Term Effects
Accuracy
Scientific Integrity

Promotion of Human Values:
Leveraged to Benefit Society
Human Values and Human Flourishing
Access to Technology

BERKMAN KLEIN CENTER
FOR INTERNET & SOCIETY
AT HARVARD UNIVERSITY



White Paper
Much as the principles documents underlying our research come from a wide variety of stakeholders in the ongoing conversation around ethical and rights-respecting AI, so too we expect a variety of readers for these materials. It is our hope that they will be useful to policymakers, academics, advocates, and technical experts. However, different groups may wish to engage with the white paper in different ways:

- Those looking for a **high-level snapshot of the current state of thinking in the governance of AI** may be best served by reviewing the data visualization (p. 8), and reading the Executive Summary (p. 4) and Human Rights section (p. 64), dipping into the discussion of themes (beginning p. 20) only where they are necessary to resolve a particular interest or question.
- Those looking to do **further research** on AI principles will likely find the discussions of the themes and principles (beginning p. 20) and Human Rights section (p. 64) most useful, and are also invited to contact the authors with requests to access the underlying data.
- Those tasked with **drafting a new set of principles** may find that the data visualization (p. 8) and discussions of the themes and principles within them (beginning p. 20) can function to offer a head start on content and approaches thereto, particularly as references to existing principles that are most likely to be useful source material.
- Those seeking closer **engagement with primary source documents** may variously find the data visualization (p. 8), timeline (p. 18), or bibliography (p. 68) to act as a helpful index.

2. Definitions and Methodology

Definition of Artificial Intelligence

The definition of artificial intelligence, or "AI", has been widely debated over the years, in part because the definition changes as technology advances.[4] In collecting our dataset, we did not exclude documents based on any particular definition of AI. Rather, we included documents that refer specifically to AI or a closely equivalent term (for example, IEEE uses "autonomous and intelligent systems").[5] In keeping with the descriptive approach we have taken in this paper, we'll share a few definitions found in our dataset. The European Commission's High-Level Expert Group on Artificial Intelligence offers a good place to start:

"Artificial intelligence (AI) systems are software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal. AI systems can either use symbolic rules or learn a numeric model, and they can also adapt their behaviour by analysing how the environment is affected by their previous actions."[6]

Aspects of this definition are echoed in those found in other documents. For example, some documents define AI as systems that take action, with autonomy, to achieve a predefined goal, and some add that these actions are generally tasks that would otherwise require human intelligence.[7]

[4] This is known as the "odd paradox" – when technologies lose their classification as "AI" because more impressive technologies take their place. *See*, Pamela McCorduck, 'Machines Who Think: A Personal Inquiry into the History and Prospects of Artificial Intelligence', 2nd ed. (Natick, MA: A. K. Peters, Ltd., 2004).

[5] 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' (2019) First Edition <https://ethicsinaction.ieee.org/>.

[6] European Commission's High-Level Expert Group on Artificial Intelligence, 'Ethics Guidelines for Trustworthy AI' (2019) p. 36 <https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai>.

[7] UK House of Lords, Select Committee on Artificial Intelligence, 'AI in the UK: Ready, Willing and Able?' (2018) Report of Session 2017-19 <https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf>; Mission assigned by the French Prime Minister, 'For a Meaningful Artificial Intelligence: Toward a French and European Strategy' (2018) <https://www.aiforhumanity.fr/pdfs/MissionVillani_Report_ENG-VF.pdf>..

Other documents define AI by the types of tasks AI systems accomplish – like “learning, reasoning, adapting, and performing tasks in ways inspired by the human mind,”[8] or by its sub-fields like knowledge-based systems, robotics, or machine learning.[9]

Definition of Relevant Documents

While all of the documents use the term “AI” or an equivalent, not all use the term “principles,” and delineating which documents on the subject of ethical or rights-respecting AI should be considered “principles” documents was a significant challenge. Our working definition was that principles are normative (in the sense that lawyers use this term) declarations about how AI generally *ought* to be developed, deployed, and governed. While the intended audience of our principles documents varies, they all endeavor to shape behavior of an audience - whether internal company principles to follow in AI development or broadly targeted principles meant to further develop societal norms about AI.

Because a number of documents employed terminology other than “principles” while otherwise conforming to this definition, we included them.[10] The concept of “ethical principles” for AI has encountered pushback both from ethicists, some of whom object to the imprecise usage of the term in this context, as well as from some human rights practitioners, who resist the recasting of fundamental human rights in this language. Rather than disaggregate AI principles from the other structures (international human rights, domestic or regional regulations, professional norms) in which they are intertwined, our research team took pains to assess principles documents in context and to flag external frameworks where relevant. In doing so, we drew inspiration from the work of Urs Gasser, Executive Director of the Berkman Klein Center for Internet & Society and Professor of Practice at Harvard Law School, whose theory on “digital constitutionalism” describes the significant role the articulation of principles by a diverse set of actors might play as part of the “proto-constitutional discourse” that leads to the crystallization of comprehensive governance norms.

Our definition of principles excluded documents that were time-bound in the sense of observations about advances made in a particular year[11] or goals to be accomplished over a particular period. It also excluded descriptive statements about AI’s risks and benefits. For example, there are numerous compelling reports that assess or comment on the ethical implications of AI, some even containing recommendations for next steps, that don’t advance a particular set of principles[12] and were thus excluded from this dataset. However, where a report included a recommendations section which did correspond to our definition, we included that section (but not the rest of the report) in our dataset,[13] and more generally, when only a certain page range from a broader document conformed to our definition, we limited our sample to those pages. The result of these choices is a narrower set of documents that we hope lends itself to side-by-side comparison, but notably excludes some significant literature.

We also excluded documents that were formulated solely as calls to a discrete further action, for example that that funding be committed, new agencies established, or additional research done on a particular topic, because they function more as a policy objective than a principle. By this same logic, we excluded national AI strategy documents that call for the creation of principles without advancing any.[14] However, where documents otherwise met our definition but contained individual principles such as calls for further research or regulation of AI (under the Accountability theme, see Section 3.2), we did include them. We also included the principle that those building and implementing AI should routinely consider the long-term effects of their work (under Professional Responsibility, see Section 3.7). Rather than constitute a discrete task, this call for further consideration functions as a principle in that it advocates that a process of reflection be built into the development of any AI system.

Finally, we excluded certain early instances of legislation or regulation which closely correspond to our definition of principles.[15] The process underlying the passage of governing law is markedly different than the one which resulted in other principles documents we were considering, and we were conscious of the fact that the goal of this project was to facilitate side-by-side comparison, and wanted to select documents that could fairly be evaluated that way. For the same reason, we excluded documents that looked at only a specific type of technology, such as facial recognition. We found that the content of principles documents was strongly affected by restrictions of technology type, and thus side-by-side comparison of these documents with others that focused on AI generally was unlikely to be maximally useful. On the other hand, we included principles documents that are sector-specific, focusing for example on the impacts of AI on the workforce or criminal justice, because they were typically similar in scope to the general documents.

[8] Information Technology Industry Council, ‘AI Policy Principles’ (2017) <https://www.itic.org/resources/AI-Policy-Principles-FullReport2.pdf>.

[9] German Federal Ministry of Education and Research, the Federal Ministry for Economic Affairs and Energy, and the Federal Ministry of Labour and Social Affairs, ‘Artificial Intelligence Strategy’ (2018) <https://www.ki-strategie-deutschland.de/home.html>; Access Now, ‘Human Rights in the Age of Artificial Intelligence’ (2018) <https://www.accessnow.org/cms/assets/uploads/2018/11/AI-and-Human-Rights.pdf>.

[10] For example, the Partnership on AI’s document is the “Tenets,” the Public Voice and European High Level Expert Group’s documents are styled as “guidelines,” the Chinese AI Industry’s document is a “Code of Conduct” and the Toronto Declaration refers to “responsibilities” in Principle 8.

[11] AI Now Institute, New York University, ‘AI Now Report 2018’ (December 2018) https://ainowinstitute.org/AI_Now_2018_Report.pdf.

[12] AI Now Institute, New York University, ‘AI Now Report 2018’ (December 2018) https://ainowinstitute.org/AI_Now_2018_Report.pdf.

[13] *See generally*, Access Now (n 9).

[14] For example, in 2017 the government of Finland published *Finland’s Age of Artificial Intelligence*, which was excluded from our dataset because it does not include principles for socially beneficial AI. *See*, Ministry of Economic Affairs and Employment of Finland, ‘Finland’s Age of Artificial Intelligence’ (2017) http://julkaisut.valtioneuvosto.fi/bitstream/handle/10024/160391/TEMrap_47_2017_verkkojulkaisu.pdf?sequence=1&isAllowed=y

[15] *See*, Treasury Board of Canada Secretariat, ‘Directive on Automated Decision-Making’ (Feb. 2019) https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592

Due to the flexibility of our definition, there remains a broad range among the documents we did include, from high-level and abstract statements of values, to more narrowly focused technical and policy recommendations. While we questioned whether this should cause us to narrow our focus still further, because the ultimate goal of this project is to provide a description of the current state of the field, we decided to retain the full range of principle types we observed in the dataset, and encourage others to dive deeper into particular categories according to their interests.

Document Search Methodology

The dataset of thirty-six documents on which this report and the associated data visualization are based was assembled using a purposive sampling method. Because a key aim of the project from the start was to create a data visualization that would facilitate side by side comparison of individual documents, it was important that the dataset be manageably sized, and also that it represent a diversity of viewpoints in terms of stakeholder, content, geography, date, and more. We also wanted to ensure that widely influential documents were well represented. For this reason, we determined that purposive sampling with the goal of maximum variation among influential documents in this very much emergent field was the most appropriate strategy.[16]

Our research process included a wide range of tools and search terms. To identify eligible documents, our team used a variety of search engines, citations from works in the field, and expertise and personal recommendations from others in the Berkman Klein Center community. Because the principles documents are not academic publications, we did not make extensive use of academic databases. General search terms included a combination of “AI” or “artificial intelligence” and “principles,” “recommendations,” “strategy,” “guideline,” and “declaration,” amongst others. We also used knowledge from our community to generate the names of organizations – companies, governments, civil society actors, etc. – might have principles documents, and then we then searched those organizations’ websites and publications.

In order to ensure that each document earned its valuable real estate in our visualization, we required that it represent the views of an organization or institution; be authored by relatively senior staff; and, in cases of multistakeholder documents, contain a breadth of involved experts. It is worth noting that some government documents are expert reports commissioned by governments rather than the work of civil servants, but all documents included in this category were officially published.

Our search methodology has some limitations. Due to the language limitations of our team, our dataset only contains documents available in English, Chinese, French, German, and Spanish. While we strove for broad geographical representation, we were unable to locate any documents from the continent of Africa, although we understand that certain African states may be currently engaged in producing AI national strategy documents which may include some form of principles. Furthermore, we recognize the possibility of network bias – because these principles documents are often shared through newsletters or mailing lists, we discovered some documents through word of mouth from those in our network. That being said, we do not purport to have a complete dataset, an admirable task which has been taken up by others.[17] Rather we have put together a selection of prominent principles documents from an array of actors.

Principle and Theme Selection Methodology

As principles documents were identified, they were reviewed in team meetings for conformity with our criteria. Those that met the criteria were assigned to an individual team member for hand coding. That team member identified the relevant pages of the document, in the case that the principles formed a sub-section of a longer document, and hand-coded all text in that section. In the initial phase, team members were actively generating the principle codes that form the basis of our database. They used the title of the principle in the document, or if no title was given or the title did not thoroughly capture the principle’s content, paraphrased the content of the principle. If an identical principle had already been entered into the database, the researcher coded the new document under that principle rather than entering a duplicate.

When the team had collected and coded approximately twenty documents, we collated the list of principles, merging close equivalents, to form a final list of forty-seven principles. We then clustered the principles, identifying ones that were closely related both in terms of their dictionary meanings (e.g. fairness and non-discrimination) as well as ones that were closely linked in the principles documents themselves (e.g. transparency and explainability). We arrived at eight total themes, each with between three and ten principles under it:

- Privacy (8 principles)
- Accountability (10 principles)
- Safety and security (4 principles)
- Transparency and explainability (8 principles)
- Fairness and non-discrimination (6 principles)
- Human control of technology (3 principles)
- Professional responsibility (5 principles)
- Promotion of human values (3 principles)

[16] For background on purposive sampling, *See* Patton, M. Q., “Qualitative evaluation and research methods” (1990) (2nd ed.). Newbury Park, CA: Sage Publications.

[17] Anna Jobin, Marcello Ienca and Effy Vayena, “The global landscape of AI ethics guidelines”, Nature Machine Intelligence (September 2019) https://doi.org/10.1038/s42256-019-0088-2; https://algorithmwatch.org/en/project/ai-ethics-guidelines-global-inventory/

We also collected data on references to human rights in each document, whether to human rights as a general concept or to specific legal instruments such as the UDHR or the ICCPR. While this data is structured similarly to the principles and themes, with individual references coded under the heading of International Human Rights, because the references appear in different contexts in different documents and we do not capture that in our coding, we do not regard it as a theme in the same way that the foregoing concepts are. See Section 4 for our observations of how the documents in our dataset engage with human rights.

Both the selection of principles that would be included in the dataset and the collation of those principles into themes were subjective, though strongly informed by content of the early documents in our dataset and the researchers' immersion in them. This has led to some frustrations about their content. For example, when we released the draft data visualization for feedback, we were frequently asked why sustainability and environmental responsibility did not appear more prominently. While the authors are sensitive to the significant impact AI is having, and will have, on the environment,[18] we did not find a concentration of related concepts in this area that would rise to the level of a theme, and as such have included the principle of "environmental responsibility" under the Accountability theme as well as discussion of AI's environmental impacts in the "leveraged to benefit society" principle under the Promotion of Human Values theme. It may be that as the conversation around AI principles continues to evolve, sustainability becomes a more prominent theme.

Following the establishment of the basic structure of principles and themes, we were conservative in the changes we made because work on the data visualization, which depended on their consistency, was already underway. We did refine the language of the principles in the dataset, for example from "Right to Appeal" to "Ability to Appeal," when many of the documents that referenced an appeal mechanism did not articulate it as a user's right. We also moved a small number of principles from one theme to another when further analysis of their contents demanded; the most prominent example of this is that "Predictability," which was included under the Accountability theme at the time our draft visualization was released in summer 2019, has been moved to the Safety and Security theme.

Because the production of the data visualization required us to minimize the number of these changes, and because our early document collection (on which the principles and themes were originally based) was biased toward documents from the U.S. and E.U., there are a small number of principles from documents – predominantly non-Western documents – that do not fit comfortably into our dataset. For example, the Japanese AI principles include a principle of fair competition which combines intranational competition law with a caution that "[e]ven if resources related to AI are concentrated in a specific country, we must not have a society where unfair data collection and infringement of sovereignty are performed under that country's dominant position."[19] We have coded this language within the "access to technology" principle under the Promotion of Human Values theme, but it does push at the edges of our definition of that principle, and is imperfectly captured by it. Had this document been part of our initial sample, its contents might have resulted in our adding to or changing the forty-seven principles we ultimately settled on.

We therefore want to remind our readers that this is a fundamentally partial and subjective approach. We view the principles and themes we have advanced herein as simply one heuristic through which to approach AI principles documents and understand their content. Other people could have made, and will make in future, other choices about which principles to include and how to group them.

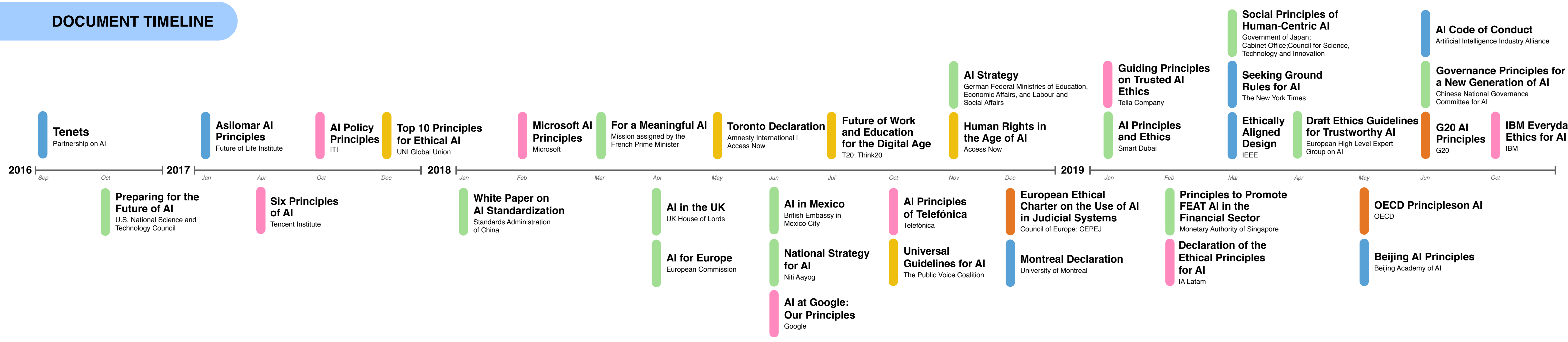
[18] Roel Dobbe and Meredith Whittaker, 'AI and Climate Change: How they're connected, and what we can do about it', AI Now Institute (2019). Retrieved from https://medium.com/@ainowinstitute.

[19] This is Principle 4.1.5. Principle of Fair Competition in Japanese Cabinet Office, Council for Science, Technology and Innovation, 'Social Principles of Human-Centric Artificial Intelligence' (2019) <https://www8.cao.go.jp/cstp/english/humancentricai.pdf>.

PRINCIPLED ARTIFICIAL INTELLIGENCE

A Map of Ethical and Rights-Based Approaches to Principles for AI

DOCUMENT TIMELINE



Tenets
Partnership on AI

Preparing for the Future of AI
U.S. National Science and Technology Council

Asilomar AI Principles
Future of Life Institute

Six Principles of AI
Tencent Institute

AI Policy Principles
ITI

Top 10 Principles for Ethical AI
UNI Global Union

White Paper on AI Standardization
Standards Administration of China

Microsoft AI Principles
Microsoft

For a Meaningful AI
Mission assigned by the French Prime Minister

AI in the UK
UK House of Lords

AI for Europe
European Commission

Toronto Declaration
Amnesty International I Access Now

AI in Mexico
British Embassy in Mexico City

National Strategy for AI
Niti Aayog

AI at Google: Our Principles
Google

Future of Work and Education for the Digital Age
T20: Think20

AI Principles of Telefónica
Telefónica

Universal Guidelines for AI
The Public Voice Coalition

AI Strategy
German Federal Ministries of Education, Economic Affairs, and Labour and Social Affairs

Human Rights in the Age of AI
Access Now

European Ethical Charter on the Use of AI in Judicial Systems
Council of Europe: CEPEJ

Montreal Declaration
University of Montreal

Guiding Principles on Trusted AI Ethics
Telia Company

AI Principles and Ethics
Smart Dubai

Principles to Promote FEAT AI in the Financial Sector
Monetary Authority of Singapore

Declaration of the Ethical Principles for AI
IA Latam

Social Principles of Human-Centric AI
Government of Japan; Cabinet Office;Council for Science, Technology and Innovation

Seeking Ground Rules for AI
The New York Times

Ethically Aligned Design
IEEE

Draft Ethics Guidelines for Trustworthy AI
European High Level Expert Group on AI

OECD Principleson AI
OECD

Beijing AI Principles
Beijing Academy of AI

AI Code of Conduct
Artificial Intelligence Industry Alliance

Governance Principles for a New Generation of AI
Chinese National Governance Committee for AI

G20 AI Principles
G20

IBM Everyday Ethics for AI
IBM

Nature of Actors

- Civil Society
- Government
- Inter-governmental Organization
- Multistakeholder
- Private Sector

BERKMAN
KLEIN CENTER
FOR INTERNET & SOCIETY
AT HARVARD UNIVERSITY

3. Themes among AI Principles

This section describes in detail our findings with respect to the eight themes, as well as the principles they contain:

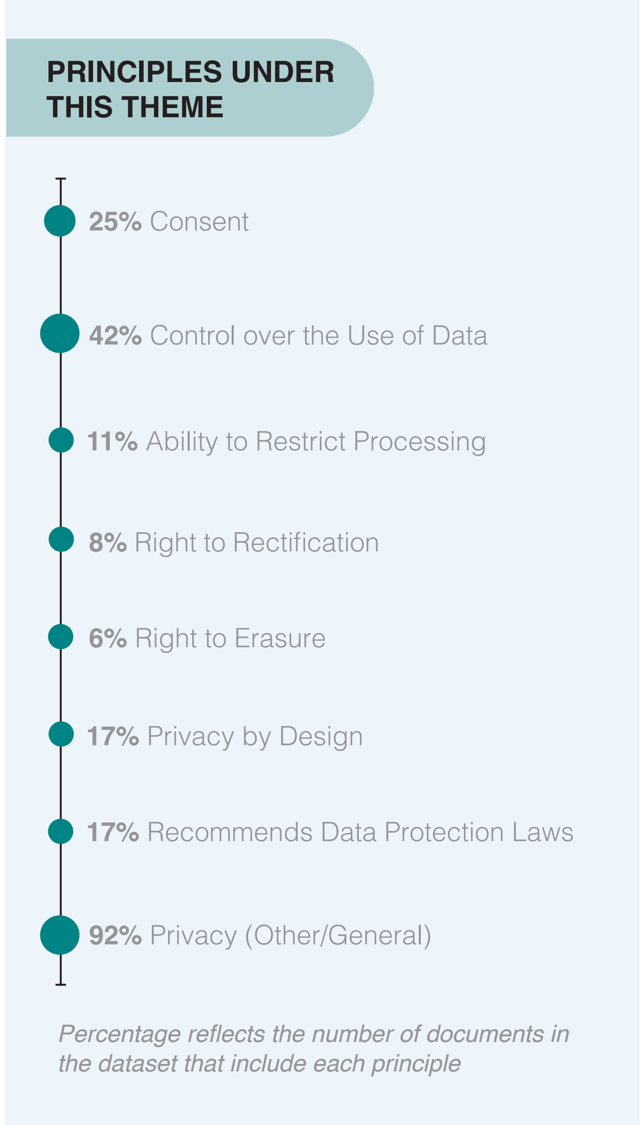
- Privacy
- Accountability
- Safety and security
- Transparency and explainability
- Fairness and non-discrimination
- Human control of technology
- Professional responsibility
- Promotion of human values

Coverage of each theme offers a view into its core features, relevance, and connection to other themes and principles. Further, we offer a detailed look at the principles under each theme, including insights generated by comparing how the principles were variously framed by the documents in our dataset.

3.1. Privacy

Privacy – enshrined in international human rights law and strengthened by a robust web of national and regional data protection laws and jurisprudence – is significantly impacted by AI technology. Fueled by vast amounts of data, AI is used in surveillance, advertising, healthcare decision-making, and a multitude of other sensitive contexts. Privacy is not only implicated in prominent implementations of AI, but also behind the scenes, in the development and training of these systems.[20] Consequently, privacy is a prominent theme[21] across the documents in our dataset, consisting of eight principles: “consent,” “control over the use of data,” “ability to restrict data processing,” “right to rectification,” “right to erasure,” “privacy by design,” “recommends data protection laws,” and “privacy (other/general).”

The General Data Protection Regulation of the European Union (GDPR) has been enormously influential in establishing safeguards for personal data protection in the current technological environment, and many of the documents in our dataset were clearly drafted with provisions of the GDPR in mind. We also see strong connections between principles under the Privacy theme and the themes of Fairness and Non-Discrimination, Safety and Security, and Professional Responsibility.



PRINCIPLES UNDER THIS THEME

25% Consent

42% Control over the Use of Data

11% Ability to Restrict Processing

8% Right to Rectification

6% Right to Erasure

17% Privacy by Design

17% Recommends Data Protection Laws

92% Privacy (Other/General)

Percentage reflects the number of documents in the dataset that include each principle

[20] Mission assigned by the French Prime Minister (n 8) p. 114 (“Yet it appears that current legislation, which focuses on the protection of the individual, is not consistent with the logic introduced by these systems [AI]—i.e. the analysis of a considerable quantity of information for the purpose of identifying hidden trends and behavior—and their effect on groups of individuals. To bridge this gap, we need to create collective rights concerning data.”) .

[21] Privacy principles are present in 97% of documents in the dataset. All of the principles written by government, private, and multistakeholder groups reference principles under the Privacy theme. Among documents sourced from civil society, only one, the Public Voice Coalition AI guidelines, did not refer to privacy.

Consent

Broadly, "consent" principles reference the notion that a person's data should not be used without their knowledge and permission. Informed consent is a closely related but more robust principle – derived from the medical field – which requires individuals be informed of risks, benefits, and alternatives. Arguably, some formulation of "consent" is a necessary component of a full realization of other principles under the Privacy theme, including "ability to restrict processing," "right to rectification," "right to erasure," and "control over the use of data."

Documents vary with respect to the depth of their description of consent, breaking into two basic categories: documents that touch lightly on it, perhaps outlining a simple notice-and-consent regime,[22] and documents that invoke informed consent specifically or even expand upon it.[23] A few documents, such as Google's AI principles and IA Latam's principles, do not go beyond defining consent as permission, but as a general matter, informed consent or otherwise non-perfunctory processes to obtain consent feature prominently in the corpus.

The boldest departures from the standard notice-and-consent model can be found in the Chinese White Paper on AI Standardization and Indian AI strategy. The Chinese document states that "the acquisition and informed consent of personal data in the context of AI should be redefined" and, among other recommendations, states "we should begin regulating the use of AI which could possibly be used to derive information which exceeds what citizens initially consented to be disclosed."[24] The Indian national strategy cautions against unknowing consent and recommends a mass-education and awareness campaign as a necessary component of implementing a consent principle in India.[25]

Control over the Use of Data

"Control over the use of data" as a principle stands for the notion that data subjects should have some degree of influence over how and why information about them is used. Certain other principles under the privacy theme, including "consent," "ability to restrict processing," "right to rectification," and "right to erasure" can be thought of as more specific instantiations of the control principle since they are mechanisms by which a data subject might exert control. Perhaps because this principle functions as a higher-level articulation, many of the documents we coded under it are light in the way of definitions for "control."

Generally, the documents in our dataset are of the perspective that an individual's ability to determine how their data is used and for what purpose should be qualified in various ways. Microsoft commits to giving consumers "*appropriate* controls so they can choose how their data is used"[26] and IEEE notes that where minors and those with diminished capacity are concerned, recourse to guardianship arrangements may be required.[27] However, several documents do contain articulations of the control principle that are more absolute. The IBM AI principles state that "Users should *always* maintain control over what data is being used and in what context."[28] On the other hand, the German AI strategy clearly states the importance of balancing and repeatedly articulates people's control over their personal data as a qualified "right." The German document suggests the use of "pseudonymized and anonymized data" as potential tools to "help strike the right balance between protecting people's right to control their personal data and harnessing the economic potential of big-data applications."[29]

There is some differentiation between the documents on the question of where control ought to reside. Some dedicate it to individuals, which is typical of current systems for data control. On the other hand, some documents would locate control in specially dedicated tools, institutions, or systems. For example, the European Commission's High-Level Expert Group describes the creation of "data protocols" and "duly qualified personnel" who would govern access to data.[30] IEEE proposes the implementation of a technology that would allow individuals to assign "an online agent" to help make "case-by-case authorization decisions as to who can process what personal data for what purpose." This technology might even be a dynamically learning AI itself – evaluating data use requests by third parties in an "autonomous and intelligent" manner.[31] Lastly, AI in the UK advocates "data trusts" that would allow individuals to "make their views heard and shape … decisions" through some combination of consultative procedures, "personal data representatives," or other mechanisms.[32]

Ability to Restrict Processing

The "ability to restrict processing" refers to the power of data subjects to have their data restricted from use in connection with AI technology. Some documents coded for this principle articulate this power as a legally enforceable right, while others stop short of doing so. For example, the Access Now report would "give people the ability to *request* that an entity stop using or limit the use of personal information."[33] Notably, Article 18 of the GDPR has legally codified this right with respect to data processing more generally, but documents within our dataset diverge in some respects from the GDPR definition.

The extent to which data subjects should be able to restrict the processing of their data is clearly in contention. For instance, the Montreal Declaration asserts that people have a "right to digital disconnection" and imposes a positive obligation

[22] *See generally* German Federal Ministry of Education and Research, the Federal Ministry for Economic Affairs and Energy, and the Federal Ministry of Labour and Social Affairs (n 9).German Federal Ministry of Education and Research, the Federal Ministry for Economic Affairs and Energy, and the Federal Ministry of Labour and Social Affairs (n 10); Google, 'AI at Google: Our Principles' (2018) <https://www.blog.google/technology/ai/ai-principles/>; Smart Dubai, 'Artificial Intelligence Principles and Ethics' (2019) <https://smartdubai.ae/initiatives/ai-principles-ethics> ;IA Latam, 'Declaración de Principios Éticos Para La IA de Latinoamérica' (2019) <http://ia-latam.com/etica-ia-latam/>; Chinese National Governance Committee for the New Generation Artificial Intelligence, led by China's Ministry of Science and Technology, 'Governance Principles for a New Generation of Artificial Intelligence: Develop Responsible Artificial Intelligence' (2019) <http://www.chinadaily.com.cn/a/201906/17/WS5d07486ba3103dbf14328ab7.html>.

[23] *See generally* Standard Administration of China and Paul Triolo, 'White Paper on Artificial Intelligence Standardization' excerpts in English published by New America (January 2018) <https://www.newamerica.org/cybersecurity-initiative/digichina/blog/translation-excerpts-chinas-white-paper-artificial-intelligence-standardization/>; Beijing Academy of Artificial Intelligence, 'Beijing AI Principles' (2019) (English translation available upon request) <https://www.baai.ac.cn/blog/beijing-ai-principles?categoryId=394>; Niti Aayog, 'National Strategy for Artificial Intelligence: #AI for All (Discussion Paper)' (2018) <https://www.niti.gov.in/writereaddata/files/document_publication/NationalStrategy-for-AI-Discussion-Paper.pdf>; IBM, 'IBM Everyday Ethics for AI' (2019) <https://www.ibm.com/watson/assets/duo/pdf/everydayethics.pdf>.

[24] Standard Administration of China and Triolo (n 24) Principle 3.3.3.

[25] Niti Aayog (n 24) p. 88.

[26] Microsoft, 'AI Principles' (2018) p. 68 <https://www.microsoft.com/en-us/ai/our-approach-to-ai> (emphasis added).

[27] IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (n 5) p. 23

[28] IBM (n 24) p. 44.

[29] German Federal Ministry of Education and Research, the Federal Ministry for Economic Affairs and Energy, and the Federal Ministry of Labour and Social Affairs (n 10) pp. 8, 16, 18, 28.

[30] European Commission's High-Level Expert Group on Artificial Intelligence (n 6) p. 17.

[31] IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (n 5) p. 23.

[32] UK House of Lords, Select Committee on Artificial Intelligence (n 8) p. 126.

[33] Access Now (n 10) p. 31.

on AI-driven systems to “explicitly offer the option to disconnect at regular intervals, without encouraging people to stay connected,”[34] and an earlier draft of the European High Level Expert Group guidelines placed a positive obligation on government data controllers to “systematically” offer an “express opt-out” to citizens.[35] However, the final version of the HLEG guidelines was far less expansive, narrowing the right to opt-out to “citizen scoring” technologies in “circumstances where … necessary to ensure compliance with fundamental rights.”[36]

Right to Rectification

The “right to rectification” refers to the right of data subjects to amend or modify information held by a data controller if it is incorrect or incomplete. As elsewhere where the word “right” is contained in the title, we only coded documents under this principle where they explicitly articulated it as a right or obligation. High-quality data contributes to safety, fairness, and accuracy in AI systems, so this principle is closely related to the themes of Fairness and Non-Discrimination and Safety and Security. Further, the “right to rectification” is closely related to the “ability to restrict processing,” insofar as they are both part of a continuum of potential responses a data subject might have in response to incorrect or incomplete information.

Rectification is not a frequently invoked principle, appearing in only three documents within our dataset. The Access Now report recommends a right to rectification closely modeled after that contained in Article 16 of the GDPR. The Singapore Monetary Authority’s AI principles place a positive obligation on firms to provide data subjects with “online data management tools” that enable individuals to review, update, and edit information for accuracy.[37] Finally, the T20 report on the future of work and education addresses this principle from a sector-specific viewpoint, describing a right held by employees and job applicants to “have access to the data held on them in the workplace and/or have means to ensure that the data is accurate and can be rectified, blocked, or erased if it is inaccurate.”[38]

Right to Erasure

The “right to erasure” refers to an enforceable right of data subjects to the removal of their personal data. Article 17 of the GDPR also contains a right to erasure, which allows data subjects to request the removal of personal data under a defined set of circumstances, and provides that the request should be evaluated by balancing rights and interests of the data holder, general public, or other relevant parties. The Access Now report models its recommendation off of Article 17, stating:

> [T]he Right to Erasure provides a pathway for deletion of a person’s personal data held by a third party entity when it is no longer necessary, the information has been misused, or the relationship between the user and the entity is terminated.[39]

However, other documents in the dataset advance a notion of the right to erasure distinct from the GDPR. Both the Chinese AI governance principles and the Beijing AI Principles include a call for “revocation mechanisms.”[40] In contrast to the Access Now articulation, the Beijing AI Principles provide for access to revocation mechanisms in “unexpected circumstances.”[41] Further, the Beijing document conditions that the data and service revocation mechanism must be “reasonable” and that practices should be in place to ensure the protection of users’ rights and interests. The version of the erasure principle in the T20 report on the future of work and education is even more narrowly tailored, and articulates a right to erasure for data on past, present, and potential employees held by employers if it is inaccurate or otherwise violates the right to privacy.[42]

Privacy by Design

“Privacy by design,” also known as data protection by design, is an obligation on AI developers and operators to integrate considerations of data privacy into the construction of an AI system and the overall lifecycle of the data. Privacy by design is codified in Article 25 of the GDPR, which stipulates data controllers must “implement appropriate technical and organisational measures...” during the design and implementation stage of data processing “to protect the rights of data subjects.”[43] Perhaps in recognition of these recent regulatory advances, IBM simply commits to adhering to national and international rights laws during the design of an AI’s data access permissions.[44]

In the private sector, privacy by design is regarded as an industry best practice, and it is under these terms that Google and Telefónica consider the principle. Google’s AI principles document does not use the phrase “privacy by design” but it does commit the company to incorporate Google’s privacy principles into the development and use of AI technologies and to “encourage architectures with privacy safeguards.”[45] Telefónica also points to its privacy policy and methodologies, stating: “In order to ensure compliance with our Privacy Policy we use a Privacy by Design methodology. When building AI systems, as with other systems, we follow Telefónica’s Security by Design approach.” ITI goes a step further, committing to “ethics by design,” a phrase that can be best understood as the integration of principles into the design of AI systems in a manner beyond what is legally required, and connects strongly with the “responsible design” principle under the Professional Responsibility theme.

Recommends Data Protection Laws

The “recommends data protection laws” principle, simply put, is that new government regulation is a necessary component of protecting privacy in the face of AI technologies. Documents produced on behalf of the governments of France, Germany,

[34] University of Montreal, ‘Montreal Declaration for a Responsible Development of Artificial Intelligence’ (2018) p. 10 (*See* Principle 3.3) <https://www.montrealdeclaration-responsibleai.com/the-declaration>.

[35] Draft European Commission’s High-Level Expert Group on Artificial Intelligence, ‘Ethics Guidelines for Trustworthy AI’ (Dec. 2018) p. 7 (*See* Principle 3.5 Citizens rights) <https://ec.europa.eu/digital-single-market/en/news/draft-ethics-guidelines-trustworthy-ai >. <https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai>.

[36] European Commission’s High-Level Expert Group on Artificial Intelligence (n 6) p. 34.

[37] Monetary Authority of Singapore, ‘Principles to Promote Fairness, Ethics, Accountability and Transparency (FEAT) in the Use of Artificial Intelligence and Data Analytics in Singapore’s Financial Sector’ (2019) p. 11 <http://www.mas.gov.sg/~/media/MAS/News%20and%20Publications/Monographs%20and%20Information%20Papers/FEAT%20Principles%20Final.pdf>.

[38] Think 20, ‘Future of Work and Education for the Digital Age’ (2018) p. 5 <https://www.g20-insights.org/wp-content/uploads/2018/07/TF1-1-11-Policy-Briefs_T20ARG_Towards-a-G20-Framework-For-Artificial-Intelligence-in-the-Workplace.pdf>.

[39] Access Now (n 10) p. 31.

[40] Chinese National Governance Committee for the New Generation Artificial Intelligence, led by China’s Ministry of Science and Technology (n 22) Principle 4.

[41] Beijing Academy of Artificial Intelligence (n 23) (*See* Principle 2.2, English translation available upon request.)

[42] Think 20 (n 39) p. 5.

[43] GDPR Art. 25 The GDPR definition and enforcement mechanism is an instructive example of privacy by design and Article 25 even specifies techniques, such as pseudonymization and data minimization, for data processors to implement.

[44] IBM (n 24) p. 44.

[45] Google (n 23) (*See* Principle 5.)

Mexico, and India each call for the development of new data privacy and data protection frameworks. These calls for regulation tend to be aspirational in their framing, with a common acknowledgement – neatly articulated in the Access Now report – that "data protection legislation can anticipate and mitigate many of the human rights risks posed by AI."[46] Other documents add that the "diverse and fast changing nature of the technology" requires a "continually updated" privacy protection regime.[47] The importance of agile regulatory frameworks is reiterated in the AI in Mexico document, which advises Mexico's National Institute for Transparency, Access to Information and Protection of Personal Data "to keep pace with innovation."[48]

The European documents that address this principle do so in the context of an already highly protective regime. The German strategy document suggests that there exists a gap in that regime, and calls for a new Workers' Data Protection Act "that would protect employees' data in the age of AI."[49] This narrow approach contrasts with the French strategy document, which critiques current legislation, and the rights framework more fundamentally, as too focused on "the protection of the individual" to adequately contend with the potential collective harms machine learning and AI systems can perpetuate. The French document calls for the creation of new "collective rights concerning data."[50] Even outside of Europe, the GDPR's influence is felt where the Indian AI strategy points towards existing practice in Europe – specifically, the GDPR and France's right to explanation for administrative algorithmic decisions – as a standard for Indian regulators to use as potential benchmarks.[51] Like the German AI strategy, the Indian AI strategy recommends establishing sector-specific regulatory frameworks to supplement a central privacy protection law.[52]

Privacy (Other/General)

Documents that were coded for the "privacy (other/general)" principle generally contain broad statements on the relevance of privacy protections to the ethical or rights-respecting development and deployment of AI. This was the single most popular principle in our dataset; nearly all of the documents in our dataset contained it.[53] Given the breadth of coverage for this principle, it's interesting to observe significant variety in the justifications for its importance. Many actors behind principles documents root the privacy principle in compliance with law, whether international human rights instruments or national or regional laws such as the GDPR, but others offer alternative rationales.

[46] Access Now (n 10) p. 30.

[47] Niti Aayog (n 24) p. 87.

[48] British Embassy in Mexico City, 'Artificial Intelligence in Mexico (La Inteligencia Artificial En México)' (2018) p. 49 <https://docs.wixstatic.com/ugd/7be025_ba24a518a53a4275af4d7ff63b4cf594.pdf>.

[49] German Federal Ministry of Education and Research, the Federal Ministry for Economic Affairs and Energy, and the Federal Ministry of Labour and Social Affairs (n 10) p. 28.

[50] Mission assigned by the French Prime Minister (n 8) p. 114.

[51] Niti Aayog (n 24) p. 87.

[52] Niti Aayog (n 24) p. 87.

[53] The three documents that did not include this principle are the Public Voice Coalition AI guidelines, the Ground Rules for AI conference paper, and the Singapore Monetary Authority's AI principles. The Public Voice Coalition AI guidelines is not coded for any principle in the Privacy theme, although in external materials such as the explanatory memorandum and references section, the organization makes it clear that privacy and data protection laws were highly influential; particularly in the framing of their "transparency" principle. *See* The Public Voice Coalition, 'Universal Guidelines for Artificial Intelligence' (2018) <https://thepublicvoice.org/ai-universal-guidelines/>.

Privacy is frequently called out as the prime example of the relevance of a rights framework to AI technology. The OECD and G20 AI principles call for "respect [for] the rule of law, human rights and democratic values," including respect for privacy.[54] The Toronto Declaration, which takes human rights as an overall framework for its approach to AI governance, also highlights the importance of privacy, stating that "States must adhere to relevant national and international laws and regulations that codify and implement human rights obligations protecting against discrimination and other related rights harms, for example data protection and privacy laws."[55] Finally, in the private sector, where AI principles most commonly take the form of internal company commitments, Telia Company engages to examine the "how we manage human rights risks and opportunities, such as privacy."[56] Other private sector actors including Microsoft, Telefónica, IA Latam, and IBM, describe respect of privacy as a legal obligation and in most cases refer to privacy as a right.

Outside of compliance, we found a wealth of other grounds for the primacy of privacy. The German AI strategy describes strong privacy standards as not only necessary from a legal and ethical standpoint but as "a competitive advantage internationally."[57] Google, and ITI describe respect of user privacy as a corporate responsibility owed to users and a business imperative.[58] The U.S. Science and Technology Council report balances consumer privacy against the value of "rich sets of data."[59] Other non-legal justifications included cybersecurity benefits,[60] alignment with public opinion,[61] and the author institution's preexisting public commitment to a set of privacy principles.[62]

[54] Organisation for Economic Co-operation and Development, 'Recommendation of the Council on Artificial Intelligence' (2019) p. 7 (*See* Principle 1.2) <https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449>; G20 Trade Ministers and Digital Economy Ministers, 'G20 Ministerial Statement on Trade and Digital Economy' (2019) p. 11 (*See* Principle 1.2) <https://www.mofa.go.jp/files/000486596.pdf>.

[55] Amnesty International, Access Now, 'Toronto Declaration: Protecting the Right to Equality and Non-Discrimination in Machine Learning Systems' (2018) p. 23 <https://www.accessnow.org/cms/assets/uploads/2018/08/The-Toronto-Declaration_ENG_08-2018.pdf>.

[56] Telia Company, 'Guiding Principles on Trusted AI Ethics' (2019) principle 3 <https://www.teliacompany.com/globalassets/telia-company/documents/about-telia-company/public-policy/2018/guiding-principles-on-trusted-ai-ethics.pdf>.

[57] German Federal Ministry of Education and Research, the Federal Ministry for Economic Affairs and Energy, and the Federal Ministry of Labour and Social Affairs (n 9) principle 16.

[58] Information Technology Industry Council (n 8) p. 1; Microsoft (n 26) p. 66.

[59] United States Executive Office of the President, National Science and Technology Council Committee on Technology, 'Preparing for the Future of Artificial Intelligence' (2016) p. 20 <https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf>.

[60] Microsoft (n 27) p. 68.

[61] IBM (n 24) p. 44.

[62] *See generally* Google (n 22); Telefónica, 'AI Principles of Telefónica' (2018) <https://www.telefonica.com/en/web/responsible-business/our-commitments/ai-principles>; Microsoft (n 26).

3.2. Accountability

On its face, the term “artificial intelligence” suggests an equivalence with human intelligence. Depending on who you ask, the age of autonomous AIs is either upon us or uncertain centuries in the future, but concerns about who will be accountable for decisions that are no longer made by humans – as well as the potentially enormous scale of this technology’s impacts on the social and natural world – likely lie behind the prevalence of the Accountability theme in our dataset.[63] Almost all documents that we analyzed mention at least one Accountability principle: “recommends adoption of new regulations,” “verifiability and replicability,” “impact assessments,” “environmental responsibility,” “evaluation and auditing requirements,” “creation of a monitoring body,” “ability to appeal,” “remedy for automated decision,” “liability and legal responsibility,” and “accountability per se.”

PRINCIPLES UNDER THIS THEME

- **36%** Verifiability and Replicability
- **53%** Impact Assessments
- **17%** Environmental Responsibility
- **47%** Evaluation and Auditing Requirement
- **17%** Creation of a Monitoring Body
- **22%** Ability to Appeal
- **11%** Remedy for Automated Decision
- **31%** Liability and Legal Responsibility
- **53%** Recommends Adoption of New Regulations
- **69%** Accountability Per Se

Percentage reflects the number of documents in the dataset that include each principle

[63] Accountability principles are present in 97% of documents in the dataset. Only one document did not mention an accountability principle. This company, Telefónica, received the highest score in the 2019 Ranking Digital Rights report, and it will be interesting to see how its ranking is impacted when, in RDR’s next report, it adds AI governance to its questionnaire. Telefónica, ‘AI Principles of Telefónica’ (October 2018).

The documents reflect diverse perspectives on the mechanisms through which accountability should be achieved. It’s possible to map the principles within the Accountability theme across the lifecycle of an AI system, in three essential stages: design (pre-deployment), monitoring (during deployment), and redress (after harm has occurred).

Design	Monitoring	Redress
Verifiability and Replicability	Evaluation and Auditing Requirements	Remedy for Automated Decision
Impact Assessment	Creation of a Monitoring Body	Liability and Legal Responsibility
Environmental Responsibility	Ability to Appeal	Recommends Adoption of New Regulations

Of course, each principle may have applicability across multiple stages as well. For example, the “verifiability and replicability” and “environmental responsibility” principles listed under the design stage in the above table will also be relevant in the monitoring and redress phases, but for optimal implementation should be accounted for when the system is designed.

The Accountability theme shows strong connections to the themes of Safety and Security, Transparency and Explainability, and Human Control of Technology.[64] Accountability principles are frequently mentioned together with the principle of transparent and explainable AI,[65] often highlighting the need for accountability as a means to gain the public’s trust[66] in AI and dissipate fears.[67]

Verifiability and Replicability

The principle of “verifiability and replicability” provides for several closely related mechanisms to ensure AI systems are functioning as they should: an AI experiment ought to “exhibit[] the same behavior when repeated under the same conditions”[68] and provide sufficient detail about its operations that it may be validated.

[64] Access Now (n 10) p. 33; Google (n 23) (*See* Principle 4.)

[65] Mission assigned by the French Prime Minister (n 8) p. 113; Amnesty International, Access Now (n 56) p. 9; UNI Global Union, ‘Top 10 Principles for Ethical Artificial Intelligence’ (2017) p. 6 <http://www.thefutureworldofwork.org/media/35420/uni_ethical_ai.pdf>; IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (n 5) pp. 29-30 (*See* Principle 6); Standard Administration of China and Triolo (n 24) (*See* Principle 3.3.1.); German Federal Ministry of Education and Research, the Federal Ministry for Economic Affairs and Energy, and the Federal Ministry of Labour and Social Affairs (n 10) p. 16; Japanese Cabinet Office, Council for Science, Technology and Innovation (n 20) p. 10 (*See* Principle 4.1.6.)

[66] German Federal Ministry of Education and Research, the Federal Ministry for Economic Affairs and Energy, and the Federal Ministry of Labour and Social Affairs (n 10) p. 16.

[67] *See e.g.*, IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (n 5) pp. 29-30 (*See* Principle 6): “We have learnt that users build a relationship with AI and start to trust it after just a few meaningful interactions. With trust, comes responsibility.“.

[68] European Commission’s High-Level Expert Group on Artificial Intelligence (n 6) p. 17.

The German AI Strategy highlights that a verifiable AI system should be able to “effectively prevent distortion, discrimination, manipulation and other forms of improper use.”[69] The development of verifiable AI systems may have institutional components along with technical ones. Institutionally, auditing institutions could “verify algorithmic decision-making in order to prevent improper use, discrimination and negative impacts on society”[70] and “new standards, including standards for validation or certification agencies on how AI systems have been verified”[71] could be developed.

Impact Assessments

The “impact assessments” principle captures both specific calls for human rights impact assessments (HRIAs) as well as more general calls for the advance identification, prevention, and mitigation of negative impacts of AI technology. One way to measure negative impacts of AI systems is to evaluate its “risks and opportunities” for human rights,[72] whether through HRIAs[73] or human rights due diligence.[74] Where HRIAs are called for, documents frequently also provide structure for their design: the Access Now report, for example, outlines that the assessment should include a consultation with relevant stakeholders “particularly any affected groups, human rights organizations, and independent human rights and AI experts.”[75] For other actors – often those less closely grounded in the daily management of technology’s human rights harms – this principle translated to calls for the assessment of “both direct and indirect harm as well as emotional, social, environmental, or other non-financial harm.”[76]

We observed that some documents use the terminology of potential *harm*[77] and others call for the identification of *risks*.[78] The emphasis, particularly among the latter category of documents, is on prevention, and impact assessments are an accountability mechanism because a sufficiently dire assessment (where risks are “too high or impossible to mitigate” [79]) should prevent an AI technology from being deployed or even developed. Some documents suggest that an AI system should only be used after evaluating its “purpose and objectives, its benefits, as well as its risks.”[80] In this context, it is particularly important that the AI system can be tested in a controlled environment and scaled-up as appropriate.[81] The Smart Dubai AI principles document calls for the use of AI systems only if they are “backed by respected and evidence-based academic research, and AI developer organizations.”[82]

[69] German Federal Ministry of Education and Research, the Federal Ministry for Economic Affairs and Energy, and the Federal Ministry of Labour and Social Affairs (n 10) p. 38.

[70] German Federal Ministry of Education and Research, the Federal Ministry for Economic Affairs and Energy, and the Federal Ministry of Labour and Social Affairs (n 10) p. 38.

[71] IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (n 5) p. 28, addressing the topic within the principle of transparency.

[72] Telia Company (n 56) (*See* Principle 3.)

[73] Access Now (n 10) p. 32; IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (n 5) p. 19; Council of Europe, European Commission For The Efficiency of Justice, ‘European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and Their Environment’ (2018) p. 8 (*See* Principle 1) <https://rm.coe.int/ethical-charter-en-for-publication-4-december-2018/16808f699c>.

[74] Amnesty International, Access Now (n 56) p. 12.

[75] Access Now (n 10) p. 34.

[76] Access Now (n 10) p. 34.

[77] Access Now (n 10) p. 34; European Commission’s High-Level Expert Group on Artificial Intelligence (n 6) p. 19.

[78] Niti Aayog (n 24) p. 87; Smart Dubai (n 23) p. 23 (*See* Principle 1.2.2.3.); Chinese National Governance Committee for the New Generation Artificial Intelligence, led by China’s Ministry of Science and Technology (n 23) (*See* Principle 8, English translation available upon request); Japanese Cabinet Office, Council for Science, Technology and Innovation (n 20) pp. 8-9 (using the term ‘harm’ within the principle of privacy protection and ‘risks’ within the principle of ensuring security; each time elaborating on impact assessment.)

[79] Amnesty International, Access Now (n 56) p. 13 (*See* para. 48.)

Environmental Responsibility

The principle of “environmental responsibility” reflects the growing recognition that AI, as a part of our human future, will necessarily interact with environmental concerns, and that those who build and implement AI technology must be accountable for its ecological impacts. The documents address environmental responsibility from two different angles.

Some documents capture this principle through an insistence that the environment should be a factor that is considered within the assessment of potential harm.[83] IA Latam’s principles, for example, stress that the impact of AI systems should not “represent a threat for our environment.”[84] Other documents go further, moving from a prohibition on negative ramifications to prescribe that AI technologies must be designed “to protect the environment, the climate and natural resources”[85] or to “promote the sustainable development of nature and society.”[86]

Evaluation and Auditing Requirement

The “evaluation and auditing requirement” principle articulates the importance of not only building technologies that are capable of being audited,[87] but also to use the learnings from evaluations to feed back into a system and to ensure that it is continually improved, “tuning AI models periodically to cater for changes to data and/or models over time.”[88]

A frequent focus is on the importance of humans in the auditing exercise, either as an auditing authority[89] or as users of AI systems who are solicited for feedback.[90] The Toronto Declaration calls upon developers to submit “systems that have a significant risk of resulting in human rights abuses to *independent* third-party audits.”[91] The T20 report on the future of work and education focuses instead on breadth of input, highlighting the need for training data and features to “be

[80] The Public Voice Coalition (n 54) (*See* Principle 5.)

[81] Organisation for Economic Co-operation and Development (n 54) p. 9 (*See* Principle 2.3); G20 Trade Ministers and Digital Economy Ministers (n 54) p. 13 (*See* Principle 2.3.)

[82] Smart Dubai (n 23) p. 22 (*See* Principle 1.2.2.1.)

[83] European Commission’s High-Level Expert Group on Artificial Intelligence (n 6) p. 19.

[84] IA Latam (n 22) (See Principle 5, English translation available upon request.)

[85] German Federal Ministry of Education and Research, the Federal Ministry for Economic Affairs and Energy, and the Federal Ministry of Labour and Social Affairs (n 10) p. 20.

[86] Beijing Academy of Artificial Intelligence (n 42) (*See* Principle 1.1 English translation available upon request.)

[87] Amnesty International, Access Now (n 54) p. 9 (*See* para. 32, particularly within the context of government acquisitions of AI systems); Mission assigned by the French Prime Minister (n 6) p. 113 (call for the development of capacities to understand and audit AI systems).

[88] Smart Dubai (n 23) p. 23 (*See* Principles 1.2.2.4 and 1.2.2.5.)

[89] Future of Life Institute, ‘Asilomar AI Principles’ (2017) p. 8 <https://futureoflife.org/ai-principles/?cn-reloaded=1>.

[90] Google (n 23) (*See* Principle 4.)

[91] Amnesty International, Access Now (n 56) p. 13 (*See* para. 47) (emphasis added).

reviewed by many eyes to identify possible flaws and to counter the 'garbage in garbage out' trap."[92]

Some, but not all, documents have drafted their "evaluation and auditing" principles to contain significant teeth. Some documents recommend the implementation of mechanisms that allow an eventual termination of use. Such a termination is recommended, in particular, if the AI systems "would violate international conventions or human rights."[93] The Access Now report suggests the development of "a failsafe to terminate acquisition, deployment, or any continued use if at any point an identified human rights violation is too high or unable to be mitigated."[94]

Creation of a Monitoring Body

The principle of "creation of a monitoring body" reflects a repeated recognition that some new organization or structure may be required to create and oversee standards and best practices in the context of AI. Visions for how these bodies may be constituted and what activities they would undertake vary.

The Ethically Aligned Design document situates the need for this new body in its pursuit to ensure that AI systems do "not infringe upon human rights, freedoms, dignity, and privacy."[95] Microsoft's AI principles suggest the creation of "internal review boards" – internal, we presume, to the company, but not to the teams that are building the technology. The Toronto Declaration stresses that any monitoring body should be independent and might include "judicial authorities when necessary."[96] The German AI strategy outlines the creation of a national AI observatory, which could also be tasked to monitor that AI systems are designed socially compatible and to develop auditing standards.[97]

Ability to Appeal

The principle of an "ability to appeal" concerns the possibility that an individual who is the subject of a decision made by an AI could challenge that decision. The ability to appeal connects with the theme of Human Control of Technology, in that it's often mentioned in connection with the principle of "right to human review of an automated decision."[98] Some documents in fact collapse the two.[99] The Access Now report calls the human in the loop an element that adds a "layer of accountability."[100]

In some individual documents, this principle is parsed more neatly, as for example in the Access Now report which explains that there should be both an ability to *challenge the use* of an AI system and an ability to *appeal a decision* that has been "informed or wholly made by an AI system."[101] The ability to appeal the use of or recommendation made by an AI system could be realized in form of a judicial review.[102] Further, some documents limit the ability to appeal only to "significant automated decisions."[103]

A subset of documents recognize as part of this principle the importance of making AI subjects aware of existing procedures to vindicate their rights[104] or to broaden the accessibility of channels for the exercise of subjects' rights.[105] In order to enable AI subjects to challenge the outcome of AI systems, the OECD and G20 AI principles suggest that the outcome of the system must be "based on plain and easy-to-understand information on the factors, and the logic that served as the basis for the prediction, recommendation or decision."[106]

Remedy for Automated Decision

The principle of "remedy for automated decision" is fundamentally a recognition that as AI technology is deployed in increasingly critical contexts, its decisions will have real consequences, and that remedies should be available just as they are for the consequences of human actions. The principle of remedy is intimately connected to the ability to appeal, since where appeal allows for the rectification of the decision itself, remedy rectifies its consequences.[107]

There is a bifurcation in many of the documents that provide for remedy between the remedial mechanisms that are appropriate for state use of AI versus those that companies should implement for private use. For example, the Toronto Declaration has separate principles for company and state action, providing that companies may "for example, creat[e] clear, independent, visible processes for redress following adverse individual or societal effects, and designat[e] roles in the entity responsible for the timely remedy of such issues"[108] whereas states should provide "reparation that, where appropriate, can involve compensation, sanctions against those responsible, and guarantees of non-repetition. This may be possible using existing laws and regulations or may require developing new ones."[109] Other documents suggest further important delineations of responsibilities, including between vendors and clients.[110]

Liability and Legal Responsibility

The principle of "liability and legal responsibility" refers to the concept that it is necessary to ensure that the individuals or entities at fault for harm

[92] Think 20 (n 39) p. 6.

[93] Partnership on AI, 'Tenets' (2016) (*See* Principle 6) <https://www.partnershiponai.org/tenets/>.

[94] Access Now (n 10) p. 33.

[95] IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (n 5) p. 19 (*See* Principle 1.)

[96] Amnesty International, Access Now (n 55) p. 10.

[97] German Federal Ministry of Education and Research, the Federal Ministry for Economic Affairs and Energy, and the Federal Ministry of Labour and Social Affairs (n 9) p. 26.

[98] UNI Global Union (n 66) p. 7; Google (n 23) (*See* Principle 4.)

[99] Think 20 (n 38) p. 8.

[100] Access Now (n 9) p. 32.

[101] Access Now (n 9) p. 33.

[102] Amnesty International, Access Now (n 55) p. 14.

[103] Smart Dubai (n 23) p. 9.

[104] Smart Dubai (n 23) p. 9.

[105] Monetary Authority of Singapore (n 38) p. 11.

[106] Organisation for Economic Co-operation and Development (n 54) p. 8 (*See* Principle 1.3); G20 Trade Ministers and Digital Economy Ministers (n 54) p. 11 (*See* Principle 1.3.)

[107] Tencent Institute (n 58) (*See* Principle 4, English translation available upon request.)

[108] Amnesty International, Access Now (n 55) p. 15 (*See* Principle 53.)

[109] Amnesty International, Access Now (n 55) p. 15 (*See* Principle 56.)

[110] Access Now (n 10) p. 35 (See para. 3.)

caused by an AI system can be held accountable. While other forms of automation and algorithmic decision making have existed for some time, emerging AI technologies can place further distance between the result of an action and the actor who caused it, raising questions about who should be held liable and under what circumstances. These principles call for reliable resolutions to those questions.

Many documents point out that existing systems may be sufficient to guarantee legal responsibility for AI harms, with actors including Microsoft and the Indian AI strategy looking to tort law and specifically negligence as a sufficient solution. Others, such as the Chinese AI Industry Code of Conduct, assert that there is additional work to be done to "[c]larify the rights and obligations of parties at each stage in research and development, design, manufacturing, operation and service of AI, to be able to promptly determine the responsible parties when harm occurs."[111]

There exists some reluctance to hold developers liable for the consequences of AI's deployment. The Chinese White Paper on AI Standardization distinguishes in its principle of liability between liability at the level of development and at the level of deployment, recommending transparency as the most appropriate accountability mechanism at the development level and suggesting the establishment of a reasonable system of liability and compensation post-deployment.[112] The Montreal Declaration makes a similar distinction, stating "[w]hen damage or harm has been inflicted by an [AI system that]… is proven to be reliable and to have been used as intended, it is not reasonable to place blame on the people involved in its development or use."[113]

Recommends Adoption of New Regulations

The "recommends adoption of new regulations" principle reflects a position that AI technology represents a significant enough departure from the status quo that new regulatory regimes are required to ensure it is built and implemented in an ethical and rights-respecting manner. Some documents that contain this principle refer to existing regulations,[114] but there is a general consensus that it is necessary to reflect on the adequacy of those frameworks.[115] Documents that contain this principle frequently express an urgent need for clarity about parties' respective responsibilities.[116] A few documents address the fact that "one regulatory approach will not fit all AI applications"[117] and emphasize the need to adopt context specific regulations, for example, regarding the use of AI for surveillance and similar activities that are likely to interfere with human rights.[118]

Among statements of this principle, we see a variety of justifications for future regulation, some of which are recognizable from other themes in our data: the regulation should ensure that the development and use of AI is safe and beneficial to society;[119] implement oversight mechanisms "in contexts that present risk of discriminatory or other rights-harming outcomes;"[120] and identify the right balance between innovation and privacy rights.[121]

There is also a common emphasis on the need for careful balancing in crafting regulation. The trade industry group ITI cautions that new regulations might "inadvertently or unnecessarily impede the responsible development and use of AI."[122] On the other hand, the OECD AI principles and G20 AI principles state that appropriate policy and regulatory frameworks can "encourage innovation and competition for trustworthy AI."[123] Many documents recognize that new laws and regulations are appropriate if lawmakers use them alongside self-regulation and existing policy tools. The AI for Europe document states that "self-regulation can provide a first set of benchmarks" but that the European Commission should "monitor developments and, if necessary, review existing legal frameworks."[124] The Standards Administration of China suggested that new regulations might be based on "universal regulatory principles"[125] that would be formulated at an international level.

Accountability Per Se

Like many of our themes, the Accountability theme contains an "accountability" principle, but in this specific case, only to those documents that explicitly use the word "accountability" or "accountable" (25 of the 36 documents) were coded under this principle. Because principles documents are frequently challenged as toothless or unenforceable, we were interested to see how documents grappled with this term specifically. In this context, documents converge on a call for developing "accountability frameworks"[126] that define the responsibility of different entities "at each stage in research and development, design, manufacturing, operation and service."[127]

Notably, a few documents emphasize that the responsibility and accountability of AI systems cannot lie with the technology itself, but should be "apportioned between those who design, develop and deploy [it]."[128] Some documents propose specific entities that should be held accountable if harm occurs, including the government,[129]

[111] Artificial Intelligence Industry Alliance, 'Artificial Intelligence Industry Code of Conduct (Consultation Version)' (2019) (*See* Principle 8, English translation available upon request) <https://www.secrss.com/articles/11099>.

[112] Standard Administration of China (n 23) (*See* Principle 3.3.2.)

[113] University of Montreal (n 35) p. 16 (*See* Principle 9.5.)

[114] Mission assigned by the French Prime Minister (n 8) p. 114 (referring to the French Data Protection Act of 1978 and the GDPR (2018.)

[115] European Commission, 'Artificial Intelligence for Europe: Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee, and the Committee of the Regions' COM (2018) p. 16 <https://ec.europa.eu/digital-single-market/en/news/communication-artificial-intelligence-europe> (Stressing, in particular, the need to reflect on "the suitability of some established rules on safety and civil law questions on liability.")

[116] UK House of Lords, Select Committee on Artificial Intelligence (n 8) p. 135 (*See* para. 56.)

[117] Information Technology Industry Council (n 8) p. 4.

[118] Access Now (n 9) p. 32.

[119] Beijing Academy of Artificial Intelligence (n 23) (*See* Preamble, English translation available upon request); Tencent Institute (n 58) (*See* Principle 18, English translation available upon request.)

[120] Amnesty International, Access Now (n 55) p. 11.

[121] British Embassy in Mexico City (n 49) p. 49.

[122] Information Technology Industry Council (n 8) p. 4.

[123] Organisation for Economic Co-operation and Development (n 54) p. 9 (*See* Principle 2.3); G20 Trade Ministers and Digital Economy Ministers (n 54) p. 13 (*See* Principle 2.3.)

[124] European Commission (n 115) p. 16.

[125] Standard Administration of China (n 23) (See Principle 3.3.1.)

[126] Information Technology Industry Council (n 8) p. 4.

[127] Artificial Intelligence Industry Alliance (n 111) (*See* Article 8, English translation available upon request.)

[128] Smart Dubai (n 22) p. 7.

[129] Access Now (n 9) p. 33.

companies and their business partners,[130] researchers, developers and users.[131] The OECD AI principles and G20 AI principles suggest that accountability should adapt to the context in which the technology is used.[132]

[130] Telia Company (n 56) (*See* Principle 5.)

[131] University of Montreal (n 34) p. 14 (*See* Principle 7.)

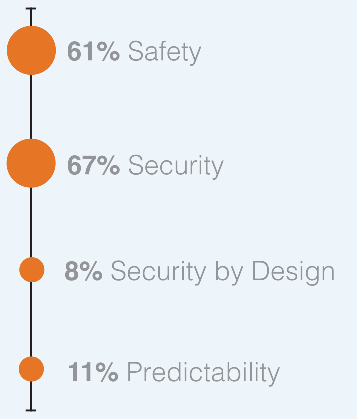
[132] Organisation for Economic Co-operation and Development (n 54) p. 8 (*See* Principle 1.5); G20 Trade Ministers and Digital Economy Ministers (n 54) p. 12 (*See* Principle 1.5.)

3.3. Safety and Security

Given early examples of AI systems' missteps[133] and the scale of harm they may cause, concerns about the safety and security of AI systems were unsurprisingly a significant theme among principles in the documents we coded.[134] There appears to be a broad consensus across different actor types on the centrality of Safety and Security, with about three-quarters of the documents addressing principles within this theme. There are four principles under it: "safety," "security," "security by design," and "predictability."

It is worth distinguishing, up front, the related concepts of safety and security. The principle of safety generally refers to proper internal functioning of an AI system and the avoidance of unintended harms. By contrast, security addresses external threats to an AI system. However, documents in our dataset often mention the two principles together, and indeed they are closely intertwined. This observation becomes particularly evident when documents use the related term "reliability":[135] a system that is reliable is safe, in that it performs as intended, and also secure, in that it is not vulnerable to being compromised by unauthorized third parties.

PRINCIPLES UNDER THIS THEME



Percentage reflects the number of documents in the dataset that include each principle

There are connections between this theme and the Accountability, Professional Responsibility, and Human Control of Technology themes. In many ways, principles under these other themes can be seen, at least partially, as implementation mechanisms for the goals articulated under Safety and Security.

[133] *See* National Transportation Safety Board Office of Public Affairs, "'Inadequate Safety Culture' Contributed to Uber Automated Test Vehicle Crash - NTSB Calls for Federal Review Process for Automated Vehicle Testing on Public Roads," (Nov. 19, 2019), <https://www.ntsb.gov/news/press-releases/Pages/NR20191119c.aspx> (describing the results of the National Transportation Safety Board's investigation in the fatal collision between an automated test vehicle operated by Uber and a pedestrian in Tempe, Arizona. Stating: "Contributing to the crash was Uber ATG's inadequate safety risk assessment procedures, ineffective oversight of the vehicle operators and a lack of adequate mechanisms for addressing operators' automation complacency – all consequences of the division's inadequate safety culture."); *see also* Cade Metz and Scott Blumenthal, "How A.I. Could be Weaponized to Spread Disinformation," *The New York Times*, (June 7, 2019), <https://www.nytimes.com/interactive/2019/06/07/technology/ai-text-disinformation.html> (discussing the disinformation threat AI driven technologies that can create "false images and sounds that are indistinguishable from the real thing" and automated text-generation systems might pose to the online information ecosystem.)

[134] Safety and Security principles are present in 81% of documents in the dataset.

[135] Microsoft (n 27) p. 61; Partnership on AI (n 94) (*See* Principle 6); Beijing Academy of Artificial Intelligence (n 24) (*See* Principle 1.4, English translation available upon request); Japanese Cabinet Office, Council for Science, Technology and Innovation (n 20) p. 10 (*See* Principle 4.1.7.); European Commission's High-Level Expert Group on Artificial Intelligence (n 6) p. 17; Think 20 (n 39) p. 7; University of Montreal (n 35) p. 8 (*See* Principle 8.3); Chinese National Governance Committee for the New Generation Artificial Intelligence, led by China's Ministry of Science and Technology (n 23) (*See* Principle 5, English translation available upon request.)

Accountability measures are key guarantors of AI safety, including verifiability[136] and the need to monitor the operation of AI systems after their deployment.[137] Individuals and organizations behind AI technology have a key role in ensuring it is designed and used in ways that are safe and secure. Safety is thus frequently mentioned in connection with the need to ensure controllability by humans.[138]

Safety

The principle of "safety" requires that an AI system be reliable and that "the system will do what it is supposed to do without harming living beings or [its] environment."[139] Articulations of this principle focus both on safety measures to be taken both before AI systems are deployed[140] and after, "throughout their operational lifetime."[141] Safety measures during development require that AI systems are "built and tested to prevent possible misuse."[142] Building systems safely means avoiding "risks of harm"[143] by assessing safety risks[144] including potential human rights violations.[145] Testing procedures should not only apply to likely scenarios, but also establish that a system "responds safely to unanticipated situations and does not evolve in unexpected ways."[146]

Testing and monitoring of AI systems should continue after deployment according to a few articulations of the "safety" principle. This is particularly relevant where the document focuses on machine learning technology, which is likely to evolve following implementation as it continues to receive input of new information. Developers of AI systems cannot always "accurately predict the risks"[147] associated with such systems ex ante. There are also safety risks associated with AI systems being implemented in ways that their creators did not anticipate, but one document suggests that designing AI that could be called safe might require the technology makes "relatively safe decisions" "even when faced with different environments in the decision-making process."[148]

Finally, two documents coded for the "safety" principle specifically call for the development of safety regulations to govern AI. One call relates specifically to the regulation of autonomous vehicles[149] and the other is more general, calling for "high standards in terms of safety and product liability"[150] within the EU. Other documents call for public awareness campaigns to promote safety.[151] For example, IEEE's Ethically Aligned Design suggests that "in the same way police officers have given public safety lectures in schools for years; in the near future they could provide workshops on safe [AI systems]."[152]

Security

The principle of "security" concerns an AI system's ability to resist external threats. Much of the language around security in our dataset is high level, but in broad terms, the documents coded here call for three specific needs to protect against security threats: the need to test the resilience of AI systems;[153] to share information on vulnerabilities[154] and cyberattacks;[155] and to protect privacy[156] and "the integrity and confidentiality of personal data."[157] With regard to the latter need, the ITI AI Policy Principles suggest that the security of data could be achieved through anonymization, de-identification, or aggregation, and they call on governments to "avoid requiring companies to transfer or provide access to technology, source code, algorithms, or encryption keys as conditions for doing business."[158] The Chinese White Paper on AI Standardization suggests that the implementation of security assurance requirements could be facilitated through a clear distribution of liability and fault between developers, product manufacturers, service providers and end users.[159]

A number of documents, concentrated in the private sector, emphasize the "integral"[160] role of security in fostering trust in AI systems.[161] The ITI AI Policy Principles state that AI technology's success depends on users' "trust that their personal and sensitive data is protected and handled appropriately."[162]

Security by Design

The "security by design" principle, as its name suggests, is related to the development of secure AI systems. The European High Level Expert Group guidelines observes that these "values-

[136] Future of Life Institute (n 90) (*See* Principle 2); Smart Dubai (n 23) p. 9.

[137] Google (n 22) (*See* Principle 3); Chinese National Governance Committee for the New Generation Artificial Intelligence, led by China's Ministry of Science and Technology (n 22) (*See* Principle 5, English translation available upon request.)

[138] Information Technology Industry Council (n 9) p. 3; Smart Dubai (n 23) p. 9. Think 20 (n 39) p. 7.

[139] European Commission's High-Level Expert Group on Artificial Intelligence (n 6) p. 17.

[140] Telia Company (n 57) (*See* Principle 6); Think 20 (n 39) p. 7; Google (n 23) (*See* Principle 3.)

[141] Future of Life Institute (n 90) (*See* Principle 2); Smart Dubai (n 23) p. 9.

[142] Telia Company (n 56) (*See* Principle 6.)

[143] Google (n 22) (*See* Principle 3.)

[144] European Commission's High-Level Expert Group on Artificial Intelligence (n 6) pp. 16-17; Organisation for Economic Co-operation and Development (n 55) p. 8 (*See* Principle 1.4); G20 Trade Ministers and Digital Economy Ministers (n 55) pp. 11-12 (*See* Principle 1.4); Smart Dubai (n 23) p. 9; Think 20 (n 39) p. 7; Japanese Cabinet Office, Council for Science, Technology and Innovation (n 20) p. 10 (*See* Principle 4.1.7.); The Public Voice Coalition (n 54) (*See* Principle 8); Beijing Academy of Artificial Intelligence (n 24) (*See* Principle 1.4, English translation available upon request.)

[145] Partnership on AI (n 94) p. 6.

[146] Think 20 (n 39) p. 7; Japanese Cabinet Office, Council for Science, Technology and Innovation (n 20) p. 9 (stating, "it is not always possible for AI to respond appropriately to rare events or deliberate attacks" and consequently arguing that society should be empowered to balance risks and benefits.)

[147] Standard Administration of China (n 23) (*See* Principle 3.3.1.)

[148] Standard Administration of China (n 23) (*See* Principle 3.3.1.)

[149] United States Executive Office of the President, National Science and Technology Council Committee on Technology (n 59) p. 17.

[150] European Commission (n 115) p. 15.

[151] Tencent Institute (n 58) (*See* Principle 16, English translation available upon request); Japanese Cabinet Office, Council for Science, Technology and Innovation (n 19) p. 9 (*See* Principle 4, stating "Society should always be aware of the balance between the benefits and risks.")

[152] IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (n 5) p. 31 (*See* Principle 5.)

[153] Google (n 22) (*See* Principle 3.)

[154] University of Montreal (n 34) p. 15 (*See* Principle 8.5.)

[155] Information Technology Industry Council (n 8) p. 4.

[156] Microsoft (n 27) p. 66; Smart Dubai (n 23) p. 9; Think 20 (n 39) p. 20; Information Technology Industry Council (n 9) p. 4; European Commission (n 116) p. 15.

[157] University of Montreal (n 34) p. 15 (*See* Principle 8.4.)

[158] Information Technology Industry Council (n 8) p. 4.

[159] Standard Administration of China (n 23) (*See* Principle 3.3.1.)

[160] Information Technology Industry Council (n 8) p. 4

[161] See IA Latam (n 23) (*See* Principle 10, English translation available upon request); Telefónica (n 63) (See Principle 4); The Public Voice Coalition (n 54) (*See* Principle 9); Telia Company (n 57) (*See* Principle 6); Google (n 23) (*See* Principle 3.)

[162] Information Technology Industry Council (n 8) p. 4.

by-design" principles may provide a link between abstract principles and specific implementation decisions.[163]

A few documents argue that existing and widely adopted security standards should apply for the development of AI systems. The German AI Strategy suggests that security standards for critical IT infrastructure should be used[164] and the Microsoft AI Principles mention that principles from other engineering disciplines of robust and fail-safe design can be valuable.[165] Similarly, the European High Level Expert Group guidelines argue for AI systems to be built with a "fallback plan" where, in the event of a problem, a system would switch its protocol "from statistical to rule-based" decision-making or require the intervention of a human before continuing.[166]

Predictability

The principle of "predictability" is concisely defined in the European High Level Expert Group guidelines, which state that for a system to be predictable, the outcome of the planning process must be consistent with the input.[167] Predictability is generally presented as a key mechanism to ensure that AI systems have not been compromised by external actors. As the German AI strategy puts it, "transparent, predictable and verifiable" AI systems may "effectively prevent distortion, discrimination, manipulation and other forms of improper use."[168] As in the "security" principle, there is an observable connection between predictable AI systems and public trust, with the Beijing AI Principles observing that improving predictability, alongside other "ethical design approaches" should help "to make the system trustworthy."[169]

[163] European Commission's High-Level Expert Group on Artificial Intelligence (n 6) p. 21.

[164] German Federal Ministry of Education and Research, the Federal Ministry for Economic Affairs and Energy, and the Federal Ministry of Labour and Social Affairs (n 10) p. 37.

[165] Microsoft (n 27) p. 64.

[166] European Commission's High-Level Expert Group on Artificial Intelligence (n 6) p. 17 (*See* Principle 1.2 Technical robustness and safety.)

[167] European Commission's High-Level Expert Group on Artificial Intelligence (n 6) p. 22.

[168] German Federal Ministry of Education and Research, the Federal Ministry for Economic Affairs and Energy, and the Federal Ministry of Labour and Social Affairs (n 10) p. 38.

[169] Beijing Academy of Artificial Intelligence (n 24) (*See* Principle 1.5, English translation available upon request.)

3.4. Transparency and Explainability

Perhaps the greatest challenge that AI poses from a governance perspective is the complexity and opacity of the technology. Not only can it be difficult to understand from a technical perspective, but early experience has already proven that it's not always clear when an AI system has been implemented in a given context, and for what task. The eight principles within the theme of Transparency and Explainability are a response to these challenges: "transparency," "explainability," "open source data and algorithms," "open government procurement," "right to information," "notification when interacting with an AI," "notification when AI makes a decision about an individual," and "regular reporting." The principles of transparency and explainability are some of the most frequently occurring individual principles in our dataset, each mentioned in approximately three-quarters of the documents.[170]

It is interesting to note a bifurcation among the principles under this theme, where some, including "explainability" and the ability to be notified when you are interacting with an AI or subject to an automated decision, are responses to entirely new governance challenges posed by the specific capabilities of current and emerging AI technologies. The rest of the principles in this theme, such as "open source data and algorithms" and "regular reporting" are well-established pillars of technology governance, now applied specifically to AI systems.

PRINCIPLES UNDER THIS THEME

- **72%** Transparency
- **78%** Explainability
- **28%** Open Source Data and Algorithms
- **3%** Open Government Procurement
- **11%** Right to Information
- **19%** Notification When AI Makes a Decision about an Individual
- **25%** Notification when Interacting with AI
- **17%** Regular Reporting

Percentage reflects the number of documents in the dataset that include each principle

[170] Transparency and Explainability principles are present in 94% of documents in the dataset. Only two documents do not include any principles under this theme. These are two government actors, the Standards Administrations of China and the report prepared by the British Embassy in Mexico City. Jeffrey Ding and Paul Triolo, "White Paper on Artificial Intelligence Standardization (Available Excerpts in English)," *New America*, January 2018, https://www.newamerica.org/cybersecurity-initiative/digichina/blog/translation-excerpts-chinas-white-paper-artificial-intelligence-standardization/; and "Artificial Intelligence in Mexico (La Inteligencia Artificial En México)" (Mexico City: British Embassy in Mexico City, June 2018), https://docs.wixstatic.com/ugd/7be025_ba24a518a53a4275af4d7ff63b4cf594.pdf.

Transparency and Explainability is connected to numerous other themes, most especially Accountability,[171] because principles within it may function as a "prerequisite for ascertaining that [such other] principles are observed."[172] It is also connected to the principle of predictability within the Safety and Security theme and to the Fairness and Non-discrimination theme.[173] The German government notes that individuals can only determine if an automated decision is biased or discriminatory if they can "examine the basis – the criteria, objectives, logic – upon which the decision was made."[174] Transparency and Explainability is a foundation for the realization of other many other principles.

Transparency

The principle of "transparency" is the assertion that AI systems should be designed and implemented in such a way that oversight of their operations are possible. The documents in the dataset vary in their suggestions about how transparency might be applied across institutions and technical systems throughout the AI lifecycle. The European High Level Expert Group guidelines note that transparency around "the data, the system, and the business models" all matter.[175] Some documents emphasize the importance of technical transparency, such as providing the relevant authorities with access to source code.[176]

Transparency throughout an AI system's life cycle means openness throughout the design, development, and deployment processes. While most documents treat transparency as binary — that is, an AI system is either transparent or it is not — several articulate the transparency principle as one that entities will strive for, with increased disclosure over time.[177] Some raise concerns about the implications of an over-broad transparency regime, which could give rise to conflicts with privacy-related principles.[178] IEEE's Ethically Aligned Design recommends the development of "new standards that describe measurable, testable levels of transparency, so that systems can be objectively assessed and levels of compliance determined."[179] Where sufficient transparency cannot be achieved, the Toronto Declaration calls upon states to "refrain from using these systems at all in high-risk contexts."[180]

Explainability

"Explainability" is defined in various ways, but is at its core about the translation of technical concepts and decision outputs into intelligible,[181] comprehensible formats suitable for evaluation. The T20 report on the future of work and education, for example, highlights the importance of "clear, complete and testable explanations of what the system is doing and why."[182] Put another way, a satisfactory explanation "should take the same form as the justification we would demand of a human making the same kind of decision."[183]

Many of the documents note that explainability is particularly important for systems that might "cause harm,"[184] have "a significant effect on individuals,"[185] or impact "a person's life, quality of life, or reputation."[186] The AI in the UK document suggests that if an AI system has a "substantial impact on an individual's life" and cannot provide "full and satisfactory explanation" for its decisions, then the system should not be deployed.[187]

The principle of explainability is closely related to the Accountability theme as well as the principle of "right to human review of automated decision" under the Human Control of Technology theme.[188] The Toronto Declaration mentions explainability as a necessary requirement to "effectively scrutinize" the impact of AI systems on "affected individuals and groups," to establish responsibilities, and to hold actors to account.[189] The European Commission's policy statement also connects explainability to the principle of nondiscrimination, as the development of understandable AI is crucial for minimizing "the risk of bias or error."[190] The need for explainability will become increasingly important as the capabilities and impact of AI systems compound.[191]

Open Source Data and Algorithms

The principle of "open source data and algorithms" is, as noted in the introduction to this theme, a familiar concept in technology governance, and it operates similarly in the context of AI as in other computer systems. The majority of documents that address it emphasize the value of the development of common algorithms[192] and open research and collaboration to support the advancement of the technology.[193] The Montreal Declaration describes this as a "socially equitable objective"[194] and the Beijing AI Principles note that open source solutions may be useful "to avoid data/platform monopolies, to share the benefits of AI development to the greatest extent, and

[171] *See*, e.g., Mission assigned by the French Prime Minister (n 7) p. 38.

[172] UNI Global Union (n 66) p. 7 (*See* Principle 1.)

[173] *See*, e.g., Council of Europe: European Commission For The Efficiency of Justice (CEPEJ), "European Ethical Charter on the Use of AI in Judicial Systems," p. 11 (*See* Principle 4); T20: Think20, "Future of Work and Education for the Digital Age", p. 7.

[174] German Federal Ministry of Education and Research, the Federal Ministry for Economic German Federal Ministry of Education and Research, the Federal Ministry for Economic Affairs and Energy, and the Federal Ministry of Labour and Social Affairs (n 10) p. 38.

[175] European Commission's High-Level Expert Group on Artificial Intelligence (n 6) p. 18.

[176] University of Montreal (n 34) (*See* Principle 5.3, stating: "[C]ode for algorithms, whether public or private, must always be accessible to the relevant public authorities and stakeholders for verification and control purposes.")

[177] Chinese National Governance Committee for the New Generation Artificial Intelligence, led by China's Ministry of Science and Technology (n 22) (*See* Principle 5, English translation available upon request); Telia Company (n 56) (*See* Principle 7); Artificial Intelligence Industry Alliance (n 111) (*See* Principle 6, English translation available upon request.)

[178] *See*, e.g., Monetary Authority of Singapore (n 37) p. 12 (*See* Principle 8.1, stating: "excessive transparency could create confusion or unintended opportunities for individuals to exploit or manipulate.")

[179] IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (n 5) p. 28.

[180] Amnesty International, Access Now (n 55) p. 9.

[181] We have coded "intelligibility," which is less common but does appear in at least three documents, as equivalent to explainability.

[182] Think 20 (n 39) p. 7.

[183] University of Montreal (n 34) p. 12 (*See* Principle 5.2.)

[184] Future of Life Institute (n 89) (*See* Principle 7); *See also*, University of Montreal (n 34) p. 12 (*See* Principle 5.2.)

[185] Smart Dubai (n 22) p. 8.

[186] University of Montreal (n 34) p. 12 (*See* Principle 5.2.)

[187] UK House of Lords, Select Committee on Artificial Intelligence (n 7) p. 40.

[188] Future of Life Institute (n 89) (*See* Principle 8.)

[189] Amnesty International, Access Now (n 55) p. 9.

[190] European Commission (n 115) p. 15.

[191] IBM (n 24) p. 28.

[192] University of Montreal (n 34) p. 13 (*See* Principle 6.7.)

[193] IA Latam (n 22) (*See* Principle 11, English translation available upon request.)

[194] University of Montreal (n 34) principle 6.7.

to promote equal development opportunities for different regions and industries."[195] Further, numerous documents also call for public and private investment in open datasets.[196]

The T20 report on the future of work and education focuses on the balance between transparency and the potential negative effect of open source policies on algorithmic innovation. One solution, they posit, is "algorithmic verifiability", which would "require companies to disclose information allowing the effect of their algorithms to be independently assessed, but not the actual code driving the algorithm."[197] Recognizing that data or algorithm disclosure is not sufficient to achieve transparency or explainability, the IEEE stresses the importance of disclosing the underlying algorithm to validation or certification agencies that can effectively serve as auditing and accountability bodies.[198]

Open Government Procurement

"Open government procurement," the requirement that governments be transparent about their use of AI systems, was only present in one document in our dataset. The Access Now report recommends that: "When a government body seeks to acquire an AI system or components thereof, procurement should be done openly and transparently according to open procurement standards. This includes publication of the purpose of the system, goals, parameters, and other information to facilitate public understanding. Procurement should include a period for public comment, and states should reach out to potentially affected groups where relevant to ensure an opportunity to input."[199]

It is notable that the Access Now report is one of the few documents in our dataset that specifically adopts a human rights framework. This principle accounts for the special duty of governments under Principle 5 of the UN Guiding Principles on Business and Human Rights to protect against human rights abuses when they contract with private businesses.

Right to Information

The "right to information" concerns the entitlement of individuals to know about various aspects of the use of, and their interaction with, AI systems. This might include "information about the personal data used in the decision-making process,"[200] "access to the factors, the logic, and techniques that produced the outcome" of an AI system,[201] and generally "how automated and machine learning decision-making processes are reached."[202]

As elsewhere where the word "right" is contained in the title of the principle, we only coded documents where they were explicitly articulated as a right or obligation. The OECD and G20 AI principles, for instance, do not call for an explicit "right to information" for users, and thus were not coded here, even though they recommend that those adversely affected by an AI system should be able to challenge it based on "easy-to-understand information on the factors, and the logic that served as the basis for the prediction, recommendation or decision."[203] One document specifically articulates the right to information as extending beyond a right to technical matter and data to the "obligation [that it] should be drawn up in plain language and be made easily accessible."[204]

Notification when AI Makes a Decision about an Individual

The definition of the principle of "notification when an AI system makes a decision about an individual" is facially fairly clear: where an AI has been employed, the person to whom it was subject should know. The AI in UK document stresses the importance of this principle to allow individuals to "experience the advantages of AI, as well as to opt out of using such products should they have concerns."[205] If people don't know when they are subject to automated decisions, they won't have the autonomy to decide whether or not they consent, or the information to reach their own conclusions about the overall value that AI provides.

In this respect, the notification principle connects to the themes of Human Control of Technology and Accountability. For example, the European Commission not only suggests that individuals should be able to opt out,[206] but also that they should be "informed on how to reach a human and how to ensure that a system's decisions can be checked or corrected,"[207] which is an important component of accountability. Access Now emphasizes the special importance of this principle when an AI system "makes a decision that impacts an individual's rights."[208]

Notification when Interacting with an AI

The principle of "notification when interacting with an AI system," a recognition of AI's increasing ability to pass the Turing test at least in limited applications, stands for the notion that humans should always be made aware when they are engaging with technology rather than directly with another person. Examples of when this principle is relevant include chatbot interactions,[209] facial recognition systems, credit scoring systems, and generally "where machine learning systems are used in the public sphere."[210]

Like "notification when an AI system makes a decision about an individual," this principle is a precondition to the actualization of other principles, including in the Accountability and Human Control of Technology themes. However, this principle is broader than the preceding one because it requires notification even in passive uses of AI systems. In the deployment of facial recognition systems, for example, the "decision"

[195] Beijing Academy of Artificial Intelligence (n 23) (*See* Principle 1.7, English translation available upon request.)

[196] Organisation for Economic Co-operation and Development (n 54) p. 8 (*See* Principle 2.1); G20 Trade Ministers and Digital Economy Ministers (n 54) p. 13 (*See* Principle 2.1.)

[197] Think 20 (n 38) p. 7.

[198] IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (n 5) p. 28 (*See* Principle 5.)

[199] Access Now (n 9) p. 32.

[200] German Federal Ministry of Education and Research, the Federal Ministry for Economic Affairs and Energy, and the Federal Ministry of Labour and Social Affairs (n 9) p. 38.

[201] The Public Voice Coalition (n 53) (*See* Principle 1.)

[202] Amnesty International, Access Now (n 55) p. 9.

[203] Organisation for Economic Co-operation and Development (n 54) p. 8 (*See* Principle 1.3); G20 Trade Ministers and Digital Economy Ministers (n 54) p. 11 (*See* Principle 1.3.)

[204] German Federal Ministry of Education and Research, the Federal Ministry for Economic Affairs and Energy, and the Federal Ministry of Labour and Social Affairs (n 10) p. 38.

[205] UK House of Lords, Select Committee on Artificial Intelligence (n 7) p. 27.

[206] European Commission's High-Level Expert Group on Artificial Intelligence (n 6) p.7.

[207] European Commission (n 115) p. 17.

[208] Access Now (n 9) p. 33.

[209] University of Montreal (n 34) p. 12 (*See* Principle 5.9.)

[210] Amnesty International, Access Now (n 55) p. 9.

principle might be interpreted to only require disclosure if an action is taken (e.g. an arrest), whereas the "interaction" principle might require notices that the facial recognition system is in use to be posted in public spaces, much like CCTV signs. Among other glosses on this principle, the European Commission notes that "consideration should be given to when users should be informed on how to reach a human"[211] and the OECD and G20 AI principles call out that that a system of notifications of AI interactions may be especially important "in the workplace."[212]

Regular Reporting

"Regular reporting" as a principle stands for the notion that organizations that implement AI systems should systematically disclose important information about their use. This might include "how outputs are reached and what actions are taken to minimize rights-harming impacts,"[213] "discovery of … operating errors, unexpected or undesirable effects, security breaches, and data leaks,"[214] or the "evaluation of the effectiveness"[215] of AI systems. The regular reporting principle can be interpreted as another implementation mechanism for transparency and explainability, and the OECD and G20 AI principles further call for governments to step in and develop internationally comparable metrics to measure AI research, development, and deployment and to gather the necessary evidence to support these claims.[216]

[211] European Commission (n 116) p. 17

[212] Organisation for Economic Co-operation and Development (n 54) p. 8 (*See* Principal 1.3); G20 Trade Ministers and Digital Economy Ministers (n 54) p. 11 (*See* Principal 1.3.)

[213] Access Now (n 9) p.33.

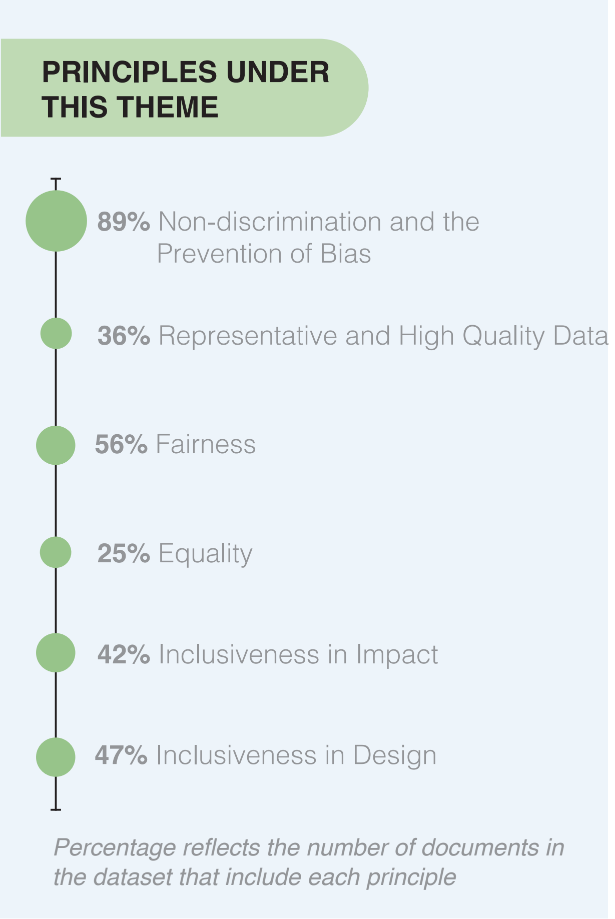
[214] University of Montreal (n 34) p. 12 (*See* Principle 5.4.)

[215] Amnesty International, Access Now (n 55) p. 49.

[216] Organisation for Economic Co-operation and Development (n 54) p. 9 (*See* Principle 2.5); G20 Trade Ministers and Digital Economy Ministers (n 54) p. 14 (*See* Principle 2.5.)

3.5. Fairness and Non-discrimination

Algorithmic bias – the systemic under- or over-prediction of probabilities for a specific population – creeps into AI systems in a myriad of ways. A system might be trained on unrepresentative, flawed, or biased data.[217] Alternatively, the predicted outcome may be an imperfect proxy for the true outcome of interest[218] or the outcome of interest may be influenced by earlier decisions that are themselves biased. As AI systems increasingly inform or dictate decisions, particularly in sensitive contexts where bias long predates their introduction such as lending, healthcare, and criminal justice, ensuring fairness and non-discrimination is imperative. Consequently, the Fairness and Non-discrimination theme is the most highly represented theme in our dataset, with every document referencing at least one of its six principles: "non-discrimination and the prevention of bias," "representative and high-quality data," "fairness," "equality," "inclusiveness in impact," and "inclusiveness in design."[219]



PRINCIPLES UNDER THIS THEME

89% Non-discrimination and the Prevention of Bias

36% Representative and High Quality Data

56% Fairness

25% Equality

42% Inclusiveness in Impact

47% Inclusiveness in Design

Percentage reflects the number of documents in the dataset that include each principle

Within this theme, many documents point to biased data – and the biased algorithms it generates – as the source of discrimination and unfairness in AI, but a few also recognize the role of human systems and institutions in perpetuating or preventing discriminatory or otherwise harmful impacts. Examples of language that focuses on the technical side of bias include the Ground Rules for AI conference paper ("[c]ompanies should strive to avoid bias in A.I. by drawing on diverse data sets")[220] and the Chinese White Paper on AI Standardization ("we should also

[217] *E.g.*, Jeffrey Dastin, "Amazon Scraps Secret AI Recruiting Tool That Showed Bias Against Women," Reuters, (Oct. 9, 2018), https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

[218] A bail decision algorithm, for example, may predict for "failure to appear" instead of flight risk to inform decisions about pretrial release. This conflates flight with other less severe causes of nonappearance (i.e. an individual may miss a court date due to inability to access transportation, childcare, or sickness) that may warrant a less punitive, lower-cost intervention than detention.

[219] Fairness and Non-discrimination principles are present in 100% of documents in the dataset.

[220] New York Times' New Work Summit, 'Seeking Ground Rules for AI' (March 2019) principle 5 <https://www.nytimes.com/2019/03/01/business/ethical-ai-recommendations.html>.

be wary of AI systems making ethically biased decisions").[221] While this concern is warranted, it points toward a narrow solution, the use of unbiased datasets, which relies on the assumption that such datasets exist. Moreover, it reflects a potentially technochauvinistic orientation – the idea that technological solutions are appropriate and adequate fixes to the deeply human problem of bias and discrimination.[222] The Toronto Declaration takes a wider view on many places bias permeates the design and deployment of AI systems:

All actors, public and private, must prevent and mitigate against discrimination risks in the design, development and application of machine learning technologies. They must also ensure that there are mechanisms allowing for access to effective remedy in place before deployment and throughout a system's lifecycle.[223]

Within the Fairness and Non-discrimination theme, we see significant connections to the Promotion of Human Values theme, with principles such as "fairness" and "equality" sometimes appearing alongside other values in lists coded under the "Human Values and Human Flourishing" principle.[224] There are also connections to the Human Control of Technology, and Accountability themes, principles under which can act as implementation mechanisms for some of the higher-level goals set by Fairness and Non-discrimination principles.

Non-discrimination and the Prevention of Bias

The "non-discrimination and the prevention of bias" principle articulates that bias in AI – in the training data, technical design choices, or the technology's deployment – should be mitigated to prevent discriminatory impacts. This principle was one of the most commonly included ones in our dataset[225] and, along with others like "fairness" and "equality" frequently operates as a high-level objective for which other principles under this theme (such as "representative and high-quality data" and "inclusiveness in design") function as implementation mechanisms.[226]

Deeper engagement with the principle of "non-discrimination and the prevention of bias" included warnings that AI is not only replicating existing patterns of bias, but also has the potential to significantly scale discrimination and to discriminate in unforeseen ways.[227] Other documents recognized that AI's great capacity for classification and differentiation could and should be proactively used to identify and address discriminatory practices in current systems.[228] The German Government commits to assessing how its current legal protections against discrimination cover – or fail to cover – AI bias, and to adapt accordingly.[229]

Representative and High Quality Data

The principle of "representative and high quality data," driven by what is colloquially referred to as the "garbage in, garbage out" problem, is defined as the use of appropriate inputs to an AI system, which relates accurately to the population of interest. The use of a dataset that is not representative leads to skewed representation of a group in the dataset compared to the actual composition of the target population, introduces bias, and reduces the accuracy of the system's eventual decisions. It is important that the data be high quality and apposite to the context in which the AI system will be deployed, because a representative dataset may nonetheless be informed by historical bias.[230] Some quality measures for data include accuracy, consistency, and validity. As the definition suggests, the documents in our dataset often directly connected this principle to the goal of mitigating the discriminatory impacts of AI.

The Montreal Declaration and the European Charter on AI in judicial systems call for representative and high quality data but state that even using the gold standard in data could be detrimental if the data are used for "deterministic analyses."[231] The Montreal Declaration's articulation of this principle warns against using data "to lock individuals into a user profile, fix their personal identity, or confine them to a filtering bubble, which would restrict and confine their possibilities for personal development."[232] Some documents, including the European Charter on AI in judicial systems, explicitly call for special protections for marginalized groups and for particularly sensitive data, defined as "alleged racial or ethnic origin, socio-economic background, political opinions, religious or philosophical beliefs, trade union membership, genetic data, biometric data, health-related data or data concerning sexual life or sexual orientation."[233]

Fairness

The "fairness" principle was defined as equitable and impartial treatment of data subjects by AI systems. We used this definition, drawn from common usage, over a technical one because articulations of fairness in the documents coded under this principle are not especially technical or overly specific in spite of the rich vein of academic research by AI and machine learning academics around competing mathematical formalizations of fairness.[234] However, Microsoft adds to its principle "AI systems should treat all people fairly" the further elaboration that "industry and academia should continue the promising work underway to develop analytical techniques to detect and address potential unfairness, like methods that systematically assess the data used to train AI systems for appropriate representativeness and document information about its origins and characteristics."[235]

[221] Standard Administration of China (n 23) (*See* Principle 3.3.2.)

[222] M. Broussard coined the term "technochauvinism" in her recent book *Artificial Unintelligence*.

[223] Amnesty International, Access Now (n 55) p.6 (*See* Principle 17.)

[224] Organisation for Economic Co-operation and Development (n 54) p. 7 (*See* Principle 1.2.); G20 Trade Ministers and Digital Economy Ministers (n 54) p. 11 (*See* Principle 1.2.)

[225] Only four documents in the dataset did not cite this principle: Asilomar AI Principles, PAI Tenets, U.S. Science and Technology Council report, and Ethically Aligned Design from the IEEE.

[226] European Commission (n 115) p. 13.

[227] Monetary Authority of Singapore (n 37) p. 6 (stating: "While the use of AIDA [Artificial Intelligence and Data Analytics] could enable analysis based on segmentation and clustering of data, this also means that differentiation between groups could take place at a greater scale and faster speed. The use of AIDA may also create the ability to identify or analyse new types of differentiation that could not previously be done. This could perpetuate cases of unjustified differentiation at a systemic level if not properly managed."); *See also*, Mission assigned by the French Prime Minister (n 7) pp. 121-122.

[228] Council of Europe, European Commission for the Efficiency of Justice (n 73) pp.9-10 (stating: "However, the use of machine learning and multidisciplinary scientific analyses to combat such discrimination should be encouraged.")

[229] German Federal Ministry of Education and Research, the Federal Ministry for Economic Affairs and Energy, and the Federal Ministry of Labour and Social Affairs (n 9) p.37.

[230] For example, a lending algorithm trained on a dataset of previously successful applicants will be "representative" of the historical applicant pool but will also replicate any past biases that informed who received a loan.

[231] Council of Europe, European Commission for the Efficiency of Justice (n 73) p. 9.

[232] University of Montreal (n 34) p.14 (*See* Principle 7.4.)

[233] European Commission's High-Level Expert Group on Artificial Intelligence (n 6).

[234] Arvind Narayanan, "Translation tutorial: 21 fairness definitions and their politics," tutorial presented at the Conference on Fairness, Accountability, and Transparency, (Feb. 23, 2018), <https://www.youtube.com/embed/jIXIuYdnyyk>

[235] Microsoft (n 27) p. 58.

There was general consensus in the documents about the importance of fairness with regard to marginalized populations. For example, the Japanese AI principles include the imperative that "all people are treated fairly without unjustified discrimination on the grounds of diverse backgrounds such as race, gender, nationality, age, political beliefs, religion, and so on."[236] Similarly, the Chinese AI Industry Code of Conduct states that "[t]he development of artificial intelligence should ensure fairness and justice, avoid bias or discrimination against specific groups or individuals, and avoid placing disadvantaged people at a more unfavorable position."[237] The European High Level Expert Group guidelines term this the "substantive dimension" of fairness, and also point to a "procedural dimension of fairness [which] entails the ability to contest and seek effective redress against decisions made by AI systems and by the humans operating them," which we coded under the "ability to appeal" principle in the Accountability theme.

Equality

The principle of "equality" stands for the idea that people, whether similarly situated or not, deserve the same opportunities and protections with the rise of AI technologies. "Equality" is similar to "fairness" but goes farther, because of fairness's focus on similar outcomes for similar inputs. As the European High Level Expert Group guidelines puts it:

> "Equality of human beings goes beyond non-discrimination, which tolerates the drawing of distinctions between dissimilar situations based on objective justifications. In an AI context, equality entails that the same rules should apply for everyone to access to information, data, knowledge, markets and a fair distribution of the value added being generated by technologies."[238]

There are essentially three different ways that equality is represented in the documents in our dataset: in terms of human rights, access to technology, and guarantees of equal opportunity through technology. In the human rights framing, the Toronto Declaration notes that AI will pose "new challenges to equality" and that "[s]tates have a duty to take proactive measures to eliminate discrimination."[239] In the access to technology framing, documents emphasize that all people deserve access to the benefits of AI technology, and that systems should be designed to facilitate that broad access.[240]

Documents that take on what we have termed the guarantees of equal opportunity framing go a bit farther in their vision for how AI systems may or should implement equality. The Montreal Declaration asserts that AI systems "must help eliminate relationships of domination between groups and people based on differences of power, wealth, or knowledge" and "must produce social and economic benefits for all by reducing social inequalities and vulnerabilities."[241] This framing makes clear the relationship between the "equality" principle and the principles of "non-discrimination and the prevention of bias" and "inclusiveness in impact."

[236] Japanese Cabinet Office, Council for Science, Technology and Innovation (n 20) p. 10.

[237] Artificial Intelligence Industry Alliance (n 111) (*See* Principle 3, English translation available upon request.)

[238] European Commission's High-Level Expert Group on Artificial Intelligence (n 6) p. 7.

[239] Amnesty International, Access Now (n 55) pp. 5, 10.

[240] European Commission's High-Level Expert Group on Artificial Intelligence (n 6) p. 18.

[241] University of Montreal (n 34) p. 13 (*See* Principles 6.2 and 6.3.)

Inclusiveness in Impact

"Inclusiveness in impact" as a principle calls for a just distribution of AI's benefits, particularly to populations that have historically been excluded. There was remarkable consensus in the language that documents employed to reflect this principle, including concepts like "shared benefits" and "empowerment":

Document	Language of principle
Asilomar AI Principles	Shared Benefit: AI technologies should benefit and empower as many people as possible.[242]
Microsoft's AI principles	Inclusiveness – AI systems should empower everyone and engage people. If we are to ensure that AI technologies benefit and empower everyone, they must incorporate and address a broad range of human needs and experiences. Inclusive design practices will help system developers understand and address potential barriers in a product or environment that could unintentionally exclude people. This means that AI systems should be designed to understand the context, needs and expectations of the people who use them.[243]
Partnership on AI Tenets	We will seek to ensure that AI technologies benefit and empower as many people as possible[244]
Smart Dubai AI principles	We will share the benefits of AI throughout society: AI should improve society, and society should be consulted in a representative fashion to inform the development of AI[245]
T20 report on the future of work and education	Benefits should be shared: AI should benefit as many people as possible. Access to AI technologies should be open to all countries. The wealth created by AI should benefit workers and society as a whole as well as the innovators.[246]
UNI Global Union's AI principles	Share the Benefits of AI Systems: AI technologies should benefit and empower as many people as possible. The economic prosperity created by AI should be distributed broadly and equally, to benefit all of humanity.[247]

The European High Level Expert Group guidelines add some detail around what "benefits" might be shared: "AI systems can contribute to wellbeing by seeking achievement of a fair, inclusive and peaceful society, by helping to increase citizen's mental autonomy, with equal distribution of economic, social and political opportunity."[248] There is a clear connection to the principles we have catalogued under the Promotion of Human Values theme, especially the principle of "leveraged to benefit society."

[242] Future of Life Institute (n 89) (*See* Principle 14.)

[243] Microsoft (n 26) p. 69.

[244] Partnership on AI (n 93) (Principle 1.)

[245] Smart Dubai (n 22) p. 11.

[246] Think 20 (n 38) p. 7

[247] UNI Global Union (n 65) p. 8 (*See* Principle 6.)

[248] European Commission's High-Level Expert Group on Artificial Intelligence (n 6) p. 9.

Inclusiveness in Design
The "inclusiveness in design" principle stands for the idea that ethical and rights-respecting AI requires more diverse participation in the development process for AI systems. This principle is expressed in two different ways. The first and more common interpretation calls for diverse AI design teams. For example, the AI for Europe document from the European Commission affirms that "More women and people of diverse backgrounds, including people with disabilities, need to be involved in the development of AI, starting from inclusive AI education and training, in order to ensure that AI is non-discriminatory and inclusive."[249] The European High Level Expert Group guidelines add that "Ideally, teams are not only diverse in terms of gender, culture, age, but also in terms of professional backgrounds and skill sets."[250]

The second interpretation holds that a broad cross-section of society should have the opportunity to weigh in on what we use AI for and in what contexts; specifically, that there should be "a genuinely diverse and inclusive social forum for discussion, to enable us to democratically determine which forms of AI are appropriate for our society."[251] The Toronto Declaration emphasizes the importance of including end users in decisions about the design and implementation of AI in order to "ensure that systems are created and used in ways that respect rights – particularly the rights of marginalised groups who are vulnerable to discrimination."[252] This interpretation is similar to the Multistakeholder Collaboration principle in our Professional Responsibility category, but it differs in that it emphasizes bringing into conversation all of society – specifically those most impacted by AI – and not just a range of professionals in, for example, industry, government, civil society organizations, and academia.

[249] European Commission (n 116) p. 13.

[250] European Commission's High-Level Expert Group on Artificial Intelligence (n 6) p. 23.

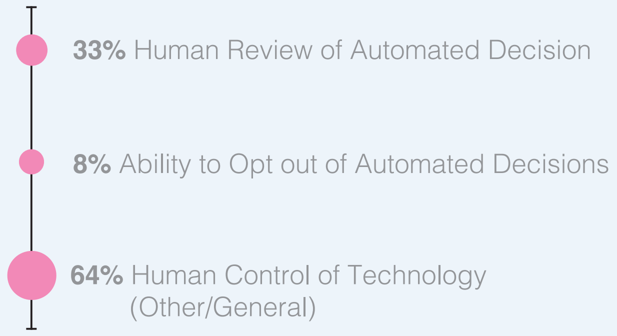
[251] Mission assigned by the French Prime Minister (n 7) p. 114.

[252] Amnesty International, Access Now (n 55) p. 6 (*See* Principle 18.)

3.6. Human Control of Technology

From prominent Silicon Valley magnates' concerns about the Singularity to popular science fiction dystopias, our society, governments, and companies alike are grappling with a potential shift in the locus of control from humans to AI systems. Thus, it is not surprising that Human Control of Technology is a strong theme among the documents in our dataset,[253] with significant representation for the three principles that fall under it: "human review of automated decision," "ability to opt out of automated decision," and "human control of technology (other/general)."

PRINCIPLES UNDER THIS THEME



Percentage reflects the number of documents in the dataset that are included each principle

There are connections between the principles in the Human Control of Technology theme and a number of other themes, because human involvement is often presented as a mechanism to accomplish those ends. Human control can facilitate objectives within the themes of Safety and Security, Transparency and Explainability, Fairness and Non-discrimination, and the Promotion of Human Values. For example, the OECD and G20 AI principles refer to human control as a "safeguard"[254] and UNI Global Union claims that transparency in both decisions and outcomes requires "the right to appeal decisions made by AI/algorithms, and having it reviewed by a human being."[255]

Human Review of Automated Decision
The principle of "human review of automated decision" stands for the idea that where AI systems are implemented, people who are subject to their decisions should be able to request and receive human review of those decisions. In contrast to other principles under this theme, the "human review of automated decision" principle is always ex post in is implementation, providing the opportunity to remedy an objectionable result. Although the documents in our dataset are situated in a variety of contexts, there is remarkable commonality between them in the articulation of this principle. The underlying rationale, when explicit, is that "Humans interacting with AI systems must be able to keep full and effective self-determination over themselves."[256]

[253] Human Control of Technology principles are present in 69% of documents in the dataset, with the Human Control of Technology (Other/General) principle most strongly represented.

[254] Organisation for Economic Co-operation and Development (n 54) p. 7 (*See* Principle 1.2); G20 Trade Ministers and Digital Economy Ministers (n 54) p. 11 (*See* Principle 1.2.)

[255] UNI Global Union (n 65) p. 7 (*See* Principle 1.)

[256] European Commission's High-Level Expert Group on Artificial Intelligence (n 6) p. 12.

The most salient differences among the documents are in the breadth of circumstances in which they suggest that human review is appropriate, and the strength of the recommendation. Many of the documents apply the principle of human review in all situations in which an AI system is used, but a handful constrain its application to situations in which the decision is "significant."[257] Further, the principles generally present human review as desirable, but two documents, the Access Now report and the Public Voice Coalition AI guidelines, articulate it as a right of data subjects. The European Charter on AI in judicial systems also contains a strong version of the human review principle, specifying that if review is requested, the case should be heard by a competent court.[258]

Ability to Opt out of Automated Decision

The "ability to opt out of automated decision" principle is defined, as its title suggests, as affording individuals the opportunity and choice not to be subject to AI systems where they are implemented. The AI in the UK document explains its relevance by saying:

> "It is important that members of the public are aware of how and when artificial intelligence is being used to make decisions about them, and what implications this will have for them personally. This clarity, and greater digital understanding, will help the public experience the advantages of AI, as well as to opt out of using such products should they have concerns."[259]

Of course, individuals interact with AI systems in numerous ways: their information may be used as training data; they may be indirectly impacted by systemic deployments of AI, and they may be personally subject to automated decisions. Perhaps because these principles are articulated with relative brevity, or perhaps because of the significant challenges in implementation, only three documents contained this principle: AI in the UK, the European High Level Expert Group guidelines, and the Smart Dubai AI principles. All documents articulated this principle as a natural corollary of the right to notification when interacting with an AI system. The latter two documents disagree about the extent of the principle's implementation, with Smart Dubai saying that entities should "consider" providing the ability to opt out "where appropriate"[260] and the European document standing for a "meaningful opportunity for human choice."[261]

Human Control of Technology (Other/General)

The "human control of technology (other/general)" principle requires that AI systems are designed and implemented with the capacity for people to intervene in their actions. This was the most commonly referenced principle[262] under the theme of Human Control of Technology, and most of the documents that included it framed it broadly, as in our definition. The Asilomar AI principles' version is illustrative: "Humans should choose how and whether to delegate decisions to AI systems, to accomplish human-chosen objectives."[263] Where the documents included a theoretical grounding for this principle, it was typically the preservation of human autonomy. For example, the Montreal Declaration states that AI systems should be built and used "respecting people's autonomy, and with the goal of increasing people's control over their lives and their surroundings."[264]

Numerous documents emphasize the importance not only of human-chosen objectives, which were included in the Asilomar principle, but the Promotion of Human Values and human quality of life.[265]Telefónica's AI Principles require that their uses of AI "be driven by value-based considerations"[266] and IA Latam's principles state that the use of AI should not only be under human control but be for the common good.[267] Others focus on the stemming the capacity of AI systems to be used to manipulate[268] or mislead people.[269]

A number of private sector principles stand out for their more granular versions of this principle, which demonstrate some connection with the theme of Professional Responsibility, because they are addressed quite directly to the developers and users of AI tools. Microsoft's AI principles include multiple steps to ensure human control, including "[e]valuation of when and how an AI system should seek human input during critical situations, and how a system controlled by AI should transfer control to a human in a manner that is meaningful and intelligible."[270] The IBM AI principles remind developers that they must identify and design for other users. The policy notes that they "may not have control over how data or a tool will be used by user, client, other external source."[271] Telia's AI principles state that the company "monitor[s] AI solutions so that we are continuously ready to intervene."[272]

Finally, emphasizing the role of people in the process in a different way, UNI Global Union asserts that AI systems must maintain "the legal status of tools, and legal persons [must] retain control over, and responsibility for, these machines at all times."[273] The Public Voice Coalition's principle of human control extends perhaps the farthest, explicitly stating that an institution has an obligation to terminate an AI system if they are no longer able to control it.[274]

[257] Smart Dubai (n 23) p. 9.

[258] Council of Europe, European Commission for the Efficiency of Justice (n 73) p. 12.

[259] UK House of Lords, Select Committee on Artificial Intelligence (n 8) p. 27.

[260] Smart Dubai (n 23) p. 26.

[261] European Commission's High-Level Expert Group on Artificial Intelligence (n 6) p. 12.

[262] 64% of documents included the "human control of technology (other/general)" principle.

[263] Future of Life Institute (n 89) (*See* Principle 16.)

[264] University of Montreal (n 34) p. 9 (*See* Principle 2.)

[265] Japanese Cabinet Office, Council for Science, Technology and Innovation (n 20) p. 4.

[266] Telefónica (n 62) (*See* Principle 3.)

[267] IA Latam (n 22) (*See* Principle 1, English translation available upon request.)

[268] European Commission's High-Level Expert Group on Artificial Intelligence (n 6) p. 12.

[269] University of Montreal (n 34) p. 9 (*See* Principle 2.)

[270] Microsoft (n 27) p. 65.

[271] IBM (n 24) p. 18.

[272] Telia Company (n 56) p. 3 (*See* Principle 4.)

[273] UNI Global Union (n 65) p. 8 (*See* Principle 4.)

[274] The Public Voice Coalition (n 53) (*See* Principle 12.)

3.7. Professional Responsibility

The theme of Professional Responsibility brings together principles that are targeted at individuals and teams who are responsible for designing, developing, or deploying AI-based products or systems. These principles reflect an understanding that the behavior of such professionals, perhaps independent of the organizations, systems, and policies that they operate within, may have a direct influence on the ethics and human rights impacts of AI. The theme of Professional Responsibility was widely represented in our dataset[275] and consists of five principles: "accuracy," "responsible design," "consideration of long-term effects," "multistakeholder collaboration," and "scientific integrity."

There are significant connections between the Professional Responsibility theme and the Accountability theme, particularly with regard to the principle of "accuracy." Articulations of the principle of "responsible design" often connect with the theme of Promotion of Human Values, and sometimes suggest Human Control of Technology as an aspect of this objective.

PRINCIPLES UNDER THIS THEME

- **19%** Accuracy
- **44%** Responsible Design
- **33%** Consideration of Long Term Effects
- **64%** Multi-stakeholder Collaboration
- **6%** Scientific Integrity

Percentage reflects the number of documents in the dataset that include each principle

Accuracy

The principle of "accuracy" is usefully defined by the European High Level Expert Group guidelines, which describe it as pertaining "to an AI's confidence and ability to correctly classify information into the correct categories, or its ability to make correct predictions, recommendations, or decisions based on data or models."[276] There is a split among the documents, with some understanding "accuracy" as a goal and others as an ongoing process.

The Google AI principles are focused narrowly on the goal of preventing the use of AI in the creation and dissemination of false information, making "accurate information readily available"[277] and the Montreal Declaration similarly avers that AI "should be designed with a view to containing [the] dissemination" of "untrustworthy information."[278] By contrast, the European High Level Expert Group guidelines are emblematic of the process-based approach, recommending that developers establish an internal definition of "accuracy" for the use case; develop a method of measurement; verify the harms caused by inaccurate predictions and measure the frequency of such predictions; and finally institute a "series of steps to increase the system's accuracy."[279] In cases when inaccurate predictions cannot be avoided, these guidelines suggest that systems indicate the likelihood of such errors.[280] Relying on a similar understanding of accuracy, the IEEE recommends operators measure the effectiveness of AI systems through methods that are "valid and accurate, as well as meaningful and actionable."[281]

The principle of accuracy is frequently referred to alongside the similar principle of "verifiability and replicability" under the Accountability theme. The Public Voice Coalition, for instance, recommends that institutions must ensure the "accuracy, reliability, and validity of decisions."[282] The two can be distinguished as "accuracy" is targeted at developers and users, promoting careful attention to detail on their part. By contrast, the principle of replicability focuses on the technology, asking whether an AI system delivers consistent results under the same conditions, facilitating post-hoc evaluation by scientists and policymakers.

Responsible Design

The principle of "responsible design" stands for the notion that individuals must be conscientious and thoughtful when engaged in the design of AI systems. Indeed, even as the phrasing of this principle might differ from document to document, there is a strong consensus that professionals are in a unique position to exert influence on the future of AI. The French AI strategy emphasizes the crucial role that researchers, engineers and developers play as "architects of our digital society."[283] This document notes that professionals play an especially important part in emerging technologies since laws and norms cannot keep pace with code and cannot solve for every negative effect that the underlying technology may bring about.[284]

The Partnership on AI Tenets prompt research and engineering communities to "remain socially responsible, and engage directly with the potential influences of AI technologies on wider society."[285] This entails, to some degree, an obligation to become informed about society, which other documents address directly. The IBM AI principles require designers and developers not only to encode values that are sensitive to different contexts but also to engage in collaboration to better recognize existing values.[286] The Tencent and Microsoft AI principles capture this idea by calling for developers to ensure that design

[275] Professional Responsibility principles are present in 78% of documents in the dataset.

[276] European Commission's High-Level Expert Group on Artificial Intelligence (n 6) p. 17.

[277] Google (n 22) (*See* Principle 1.)

[278] University of Montreal (n 34) p.9 (*See* Principle 2.5.)

[279] European Commission's High-Level Expert Group on Artificial Intelligence (n 6) p. 17.

[280] European Commission's High-Level Expert Group on Artificial Intelligence (n 6) p. 17.

[281] IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (n 5) p. 25 (*See* Principle 4.)

[282] The Public Voice Coalition (n 53) (*See* Principle 6.)

[283] Mission assigned by the French Prime Minister (n 8) p. 114.

[284] Mission assigned by the French Prime Minister (n 8) p. 114.

[285] Partnership on AI (n 93) (*See* Principle 6.)

[286] IBM (n 24) p. 22.

is “aligned with human norms in reality”[287] and to involve domain experts in the design and deployment of AI systems.[288] We note a rare interaction among the documents when the Indian AI strategy recommends that evolving best practices such as the recommendations by the Global Initiative on Ethics of Autonomous and Intelligent Systems by IEEE be incorporated in the design of AI systems.[289]

Consideration of Long Term Effects

The principle of “consideration of long term effects” is characterized by deliberate attention to the likely impacts, particularly distant future impacts, of an AI technology during the design and implementation process. The documents that address this principle largely view the potential long-term effects of AI in a pluralistic manner. For instance, the German AI strategy highlights that AI is a global development and policymakers will need to “think and act globally” while considering its impact during the development stage[290]; and the Asilomar principles recognize that highly-developed AI must be for the benefit of all of humanity and not any one sub-group.[291] The Montreal Declaration recommends that professionals must anticipate the increasing risk of AI being misused in the future and incorporate mechanisms to mitigate that risk.[292]

Some of the documents base their articulations of this principle on the premise that AI capabilities in the future may be vastly advanced compared to the technology we know today. The Beijing AI principles recommend that research on potential risks arising out of augmented intelligence, artificial general intelligence and superintelligence be encouraged.[293] These documents take the position that possibility of catastrophic or existential risks arising out of AI systems in the future cannot not be ruled out and professionals must work towards avoiding or mitigating such impacts.[294]

Multistakeholder Collaboration

“Multistakeholder collaboration” is defined as encouraging or requiring that designers and users of AI systems consult relevant stakeholder groups while developing and managing the use of AI applications. This was the most commonly included of the principles under Professional Responsibility.[295] Broadly, the documents reflect either a tool-specific or a general policy vision for multistakeholderism.

The IBM AI principles are emblematic of a tool-specific vision, specifying that developers should try to consult with policymakers and academics as they build AI systems to bring in different perspectives.[296] Additionally, the principles recommend that a feedback loop or open dialogue be established with users allowing them to highlight biases or other on-ground challenges that the system might bring about once deployed.[297] The Toronto Declaration calls for meaningful consultation with users and especially marginalized groups during the design and application of machine learning systems.[298] Access Now also suggests that human rights organizations and independent human rights and AI experts be included during such consultations.[299]

Documents that espouse a general policy function for multistakeholderism call for collaboration across the globe, rather than around any particular tool. Participants may be drawn from universities, research institutions, industry, policymaking, and the public at large to examine AI developments across sectors and use cases. The Japanese and Chinese AI strategies, for instance, push for international cooperation on AI research and use, to build a “non-regulatory, non-binding” framework.[300] This interpretation of multistakeholderism is focused on the utility of building a normative consensus on the governance of AI technologies. This vision is also seen as a policy vehicle through which governments can educate and train their populations to ensure an easy transition and safety as labor markets continue to modernize.[301]

Scientific Integrity

The principle of “scientific integrity” means that those who build and implement AI systems should be guided by established professional values and practices. Interestingly, both documents that include this relatively little-mentioned principle are organizations driven at least in significant part by engineers and technical experts. Google’s AI principles recognize scientific method and excellence as the bedrock for technological innovation, including AI. The company makes a commitment to honor “open inquiry, intellectual rigor, integrity, and collaboration” in its endeavors.[302] The IEEE acknowledges the idea of scientific rigor in its call for creators of AI systems to define metrics, make them accessible, and measure systems.[303]

[287] Tencent Institute (n 58) (*See* Principle 12, English translation available upon request.)

[288] Microsoft (n 27) p. 65.

[289] Niti Aayog (n 24) p. 87.

[290] German Federal Ministry of Education and Research, the Federal Ministry for Economic Affairs and Energy, and the Federal Ministry of Labour and Social Affairs (n 10) p. 40.

[291] Future of Life Institute (n 89) (*See* Principle 23.)

[292] University of Montreal (n 34) p. 15 (See Principle 8.)

[293] Beijing Academy of Artificial Intelligence (n 23) (*See* Principle 3.5, English translation available upon request.)

[294] Beijing Academy of Artificial Intelligence (n 23) (*See* Principle 3.5, English translation available upon request.)

[295] 64% of documents included it in one form of another.

[296] IBM (n 24) p. 24.

[297] IBM (n 24) p. 36.

[298] Amnesty International, Access Now (n 56) p. 6.

[299] Access Now (n 10) p. 34.

[300] Japanese Cabinet Office, Council for Science, Technology and Innovation (n 20) p. 6.

[301] *See* e.g., Organisation for Economic Co-operation and Development (n 54) p. 9 (*See* Principle 2.4.); G20 Trade Ministers and Digital Economy Ministers (n 54) p. 13 (*See* Principle 2.4.)

[302] Google (n 22) (*See* Principle 6.)

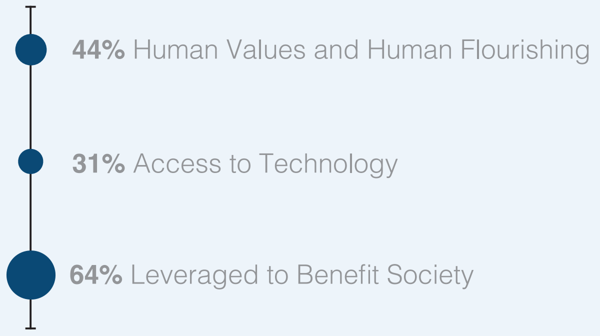
[303] IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (n 5) p. 25.

3.8. Promotion of Human Values

With the potential of AI to act as a force multiplier for any system in which it is employed, the Promotion of Human Values is a key element of ethical and rights-respecting AI.[304] The principles under this theme recognize that the ends to which AI is devoted, and the means by which it is implemented, should correspond with and be strongly influenced by social norms. As AI's use becomes more prevalent and the power of the technology increases, particularly if we begin to approach artificial general intelligence, the imposition of human priorities and judgment on AI is especially crucial. The Promotion of Human Values category consists of three principles: "human values and human flourishing," "access to technology," and "leveraged to benefit society."

While principles under this theme were coded distinctly from explicit references to human rights and international instruments of human rights law, there is a strong and clear connection. References to human values and human rights were often adjacent to one another, and where the documents provided more specific articulations of human values, they were are largely congruous with existing guarantees found in international human rights law. Moreover, principles that refer to human values often include explicit references to fundamental human rights or international human rights, or mention concepts from human rights frameworks and jurisprudence such as human dignity or autonomy. The OECD and G20 AI principles also add "internationally recognized labor rights" to this list.[305]

PRINCIPLES UNDER THIS THEME



Percentage reflects the number of documents in the dataset that include each principle

There is also an overlap between articulations of the Promotion of Human Values and social, economic, or environmental concepts that are outside the boundaries of political and civil rights,[306] including among documents coded under the principle of AI "leveraged to benefit society." Principle 3, "Make AI Serve People and Planet," from the UNI Global Union's AI principles, is emblematic, calling for: "throughout their entire operational process, AI systems [to] remain compatible and increase the principles of human dignity, integrity, freedom, privacy and cultural and gender diversity, as well as … fundamental human rights. In addition, AI systems must protect and even improve our planet's ecosystems and biodiversity."[307]

Human Values and Human Flourishing

The principle of "human values and human flourishing" is defined as the development and use of AI with reference to prevailing social norms, core cultural beliefs, and humanity's best interests. As the Chinese AI Governance Principles put it, this principle means that AI should "serve the progress of human civilization."[308] This is the broadest of the three principles in the Promotion of Human Values theme and is mentioned in 44 percent of documents. Most documents do not delve especially deeply into what they intend by "human values" beyond references to concepts like self-determination,[309] but the Montreal Declaration contains a somewhat idiosyncratic list, calling for AI systems that "permit the growth of the well-being of all sentient beings" by, inter alia, "help[ing] individuals improve their living conditions, their health, and their working conditions, … allow[ing] people to exercise their mental and physical capacities [and]… not contribut[ing] to increasing stress, anxiety, or a sense of being harassed by one's digital environment."[310]

Many of the documents that refer to the theme of "human values and human flourishing" are concerned with how the societal impacts of AI can be managed through AI system design. Tencent's AI principles state that "The R&D of artificial intelligence should respect human dignity and protect human rights and freedoms."[311] The Smart Dubai AI principles says we should "give AI systems human values and make them beneficial to society,"[312] suggesting that it is possible to build AI systems that have human values embedded in their code.[313] However, most, if not all, of these documents also acknowledge that human values will also need to be promoted in the implementation of AI systems and "throughout the AI system lifecycle."[314]

Access to Technology

The "access to technology" principle represents statements that the broad availability of AI technology, and the benefits thereof, is a vital element of ethical and rights-respecting AI. Given the significant transformational potential of AI, documents that include this principle worry that AI might contribute to the growth of inequality. The ITI AI Policy Principles, a private sector document, focus on the economic aspect, stating that "if the value [created by AI] favors only certain incumbent entities, there is a risk of exacerbating existing wage, income, and wealth gaps."[315] At least one civil society document shares this concern: the T20 report on the future of work and education

[304] Promotion of Human Values principles are present in 69% of documents in the dataset.

[305] Organisation for Economic Co-operation and Development (n 54) p. 7 (*See* Principle 1.2.); G20 Trade Ministers and Digital Economy Ministers (n 54) p. 11 (*See* Principle 1.2.)

[306] *See generally*, Future of Life Institute (n 89); European Commission (n 115).

[307] UNI Global Union (n 66) p. 7 (*See* Principle 3.)

[308] Chinese National Governance Committee for the New Generation Artificial Intelligence, led by China's Ministry of Science and Technology (n 22) (*See* Principle 1, English translation available upon request.)

[309] Think 20 (n 38) p. 7.

[310] University of Montreal (n 34) p. 8 (Principle 1.)

[311] Tencent Institute (n 58) (*See* Principle 1, English translation available upon request.)

[312] Smart Dubai (n 22) p. 10.

[313] One document from our draft dataset that is no longer included in the final version, SAGE's The Ethics of Code: Developing AI for Business with Five Core Principles has a similar Principle as found in the Smart Dubai document, stating in Principle 3: "…Reinforcement learning measures should be built not just based on what AI or robots do to achieve an outcome, but also on how AI and robots align with human values to accomplish that particular result."

[314] Organisation for Economic Co-operation and Development (n 54) p. 7 (See Principle 1.2.); G20 Trade Ministers and Digital Economy Ministers (n 54) p. 11 (*See* Principle 1.2.)

[315] Information Technology Industry Council (n 9) p. 5 (*See* "Democratizing Access and Creating Equality of Opportunity.")

avers that “The wealth created by AI should benefit workers and society as a whole as well as the innovators.”[316] The Japanese AI principles, while acknowledging the economic dimension of this issue (observing that “AI should not generate a situation where wealth and social influence are unfairly biased towards certain stakeholders”[317]), emphasize the sociopolitical dimensions of inequality, including the potential that AI may unfairly benefit certain states or regions as well as contribute to “a digital divide with so-called ‘information poor’ or ‘technology poor’ people left behind.”[318]

Some versions of the “access to technology” principle are premised on the notion that broad access to AI technology itself, as well as the education necessary to use and understand it, is the priority. The Chinese AI governance principles provide that “Stakeholders of AI systems should be able to receive education and training to help them adapt to the impact of AI development in psychological, emotional and technical aspects.”[319] The ITI AI Policy Principles focus on educating and training people who have traditionally been marginalized by or excluded from technological innovation, calling for the “diversification and broadening of access to the resources necessary for AI development and use, such as computing resources, education, and training.”[320] Two documents, Microsoft’s AI Principles and the European High Level Expert Group guidelines, go beyond this to reflect a vision for “[a]ccessibility to this technology for persons with disabilities,”[321] noting that in some cases “AI-enabled services… are already empowering those with hearing, visual and other impairments.”[322]

Leveraged to Benefit Society

The principle that AI be “leveraged to benefit society” stands for the notion that AI systems should be employed in service of public-spirited goals. The documents vary in the specificity with which they articulate goals. Where they are specific in the goals they list, they may include social, political, and economic factors. Examples of beneficial ends in the European High Level Expert Group guidelines include: “Respect for human dignity... Freedom of the individual... Respect for democracy, justice and the rule of law... Equality, non-discrimination and solidarity - including the rights of persons at risk of exclusion... Citizens’ rights… including the right to vote, the right to good administration or access to public documents, and the right to petition the administration.”[323] The High Level Expert Group and the German AI strategy were the two documents to explicitly include the environment and sustainable development as factors in their determination of AI that is “leveraged to benefit society.”[324]

There is a notable trend among the documents that include this principle to designate it as a *precondition* for AI development and use. IEEE’s Ethically Aligned Design document uses strong language to assert that it is not enough for AI systems to be profitable, safe, and legal; they must also include human well-being as a “primary success criterion for development.”[325] Google’s AI principles contain a similar concept, stating that the company “will proceed [with the development of AI technology] where we believe that the overall likely benefits substantially exceed the foreseeable risks and downsides” after taking “into account a broad range of social and economic factors.”[326]

[316] Think 20 (n 38) p. 7.

[317] Japanese Cabinet Office, Council for Science, Technology and Innovation (n 20) p. 9.

[318] Japanese Cabinet Office, Council for Science, Technology and Innovation (n 20) p. 7.

[319] Chinese National Governance Committee for the New Generation Artificial Intelligence, led by China’s Ministry of Science and Technology (n 22) (*See* Principle 2.3., English translation available upon request.)

[320] Information Technology Industry Council (n 9) p. 5 (*See* “Democratizing Access and Creating Equality of Opportunity.”)

[321] European Commission’s High-Level Expert Group on Artificial Intelligence (n 6) p. 18.

[322] Microsoft (n 27) p. 70.

[323] European Commission’s High-Level Expert Group on Artificial Intelligence (n 6) p. 11 (*See* Principle 2.1.)

[324] European Commission’s High-Level Expert Group on Artificial Intelligence (n 5) p. 32 (*See* “Example of Trustworthy AI”); German Federal Ministry of Education and Research, the Federal Ministry for Economic Affairs and Energy, and the Federal Ministry of Labour and Social Affairs (n 9) p. 9.

[325] IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (n 5) pp. 21-22 (*See* Principle 2.)

[326] Google (n 23) (*See* Principle 1.)

4. International Human Rights

In recent years, the human rights community has become more engaged with digital rights, and with the impacts of AI technology in particular. Even outside of human rights specialists, there has been an increasing appreciation for the relevance of international human rights law and standards to the governance of artificial intelligence.[327] To an area of technology governance that is slippery and fast-moving, human rights law offers an appealingly well-established core set of concepts, against which emerging technologies can be judged. To the broad guarantees of human rights law, principles documents offer a tailored vision of the specific – and in some cases potentially novel – concerns that AI raises.

Accordingly, when coding the principles documents in our dataset, we also made observations on each document's references to human rights, whether as a general concept or specific human-rights related documents such as the Universal Declaration of Human Rights, International Covenant on Civil and Political Rights, the United Nations Guiding Principles on Business & Human Rights and the United Nations Sustainable Development Goals. Twenty-three of the documents in our dataset (64%) made a reference of this kind. We also noted when documents stated explicitly that they had employed a human rights framework, and five of the thirty-six documents (14%) did so.

Given the increasing visibility of AI in the human rights community and the apparent increasing interest in human rights among those invested in AI governance, we had expected that the data might reveal a trend toward increasing emphasis on human rights in AI principles documents. However, our dataset was small enough, and the timespan sufficiently compressed, that no such trend is apparent.

As illustrated in the table below, private sector and civil society documents were most likely to reference human rights. At the outset of our research, we had expected that principles documents from the private sector would be less likely to refer to human rights and government documents more likely. Among the principles documents we looked at – admittedly not designed to be a complete or even representative sample – we were wrong. The actor type with the single greatest proportion of human rights references were the documents from the private sector; only one omitted a reference to human rights. By contrast, less than half of documents authored by or on behalf of government actors did contain some reference to human rights.[328]

[327] Filippo A. Raso, Hannah Hilligoss, and Vivek Krishnamurthy, 'Artificial Intelligence & Human Rights: Opportunities & Risks', Berkman Klein Center (September 25, 2018) https://cyber.harvard.edu/publication/2018/artificial-intelligence-human-rights.

[328] The government documents were from Europe (France, Germany, European Commission (both documents)), China and Japan.

Nature of actor	Number of documents	Number with any reference to human rights	
Civil society	5	4	80%
Government	13	6	46%
Intergovernmental organization	3	2	67%
Multistakeholder initiative	7	4	57%
Private sector	8	7	88%
Total	**36**	**23**	**64%**

There are multiple possible explanations for this. It may be that the agencies or individuals in government who have been tasked with drafting and contributing to principles documents were not selected for their expertise with human rights law, or it may be that national laws, such as the GDPR, are perceived as more relevant.

The documents also exhibit significant variation in the degree to which they are permeated by human rights law, with some using it as the framework of the whole document (denoted by a star in the data visualization), and others merely mentioning it in passing (denoted by a diamond). Using a human rights framework means that the document uses human rights as a basis for further ethical principle for the development and use of AI systems. Only five documents use a human rights framework. Three are civil society documents and two are government documents from the EU: Access Now report, AI for Europe, European High Level Expert Group guidelines, Public Voice Coalition AI guidelines, and Toronto Declaration.

5. Conclusion

The eight themes that surfaced in this research – Privacy, Accountability, Safety and Security, Transparency and Explainability, Fairness and Non-discrimination, Human Control of Technology, Professional Responsibility, and Promotion of Human Values – offer at least some view into the foundational requirements for AI that is ethical and respectful of human rights. However, there's a wide and thorny gap between the articulation of these high-level concepts and their actual achievement in the real world. While it is the intent of this white paper and the accompanying data visualization to provide a high-level overview, there remains more work to be done, and we close with some reflections on productive possible avenues.

In the first place, our discussion of the forty-seven principles we catalogued should make clear that while there are certainly points of convergence, by no means is there unanimity. The landscape of AI ethics is burgeoning, and if calls for increased access to technology (see Section 3.8) and multistakeholder participation (see Section 3.7) are heeded, it's likely to become yet more diverse. It would be compelling to have closer studies of the variation within the themes we uncovered, including additional mapping projects that might illustrate narrower or different versions of the themes with regard to particular geographies or stakeholder groups. It would also be interesting to look at principles geared toward specific applications of AI, such as facial recognition or autonomous vehicles.

Within topics like "fairness," the varying definitions and visions represented by the principles documents in our dataset layer on top of an existing academic literature,[329] but also on existing domestic and international legal regimes which have long interpreted these and similar concepts. Litigation over the harmful consequences of AI technology is still nascent, with just a handful of cases having been brought. Similarly, only a few jurisdictions have adopted regulations concerning AI, although certainly many of the documents in our dataset anticipate, and even explicitly call for (see Sections 3.1 and 3.2), such actions. Tracking how principles documents engage with and influence how liability for AI-related damages is apportioned by courts, legislatures, and administrative bodies, will be important.

There will be a rich vein for further scholarship on ethical and rights-respecting AI for some time, as the norms we attempt to trace remain actively in development. What constitutes "AI for good" is being negotiated both through top-down efforts such as dialogues at the intergovernmental level, as well as bottom-up, among people most impacted by the deployment of AI technology, and the organizations who represent their interests. That there are core themes to these conversations even now is due to the hard work of the many individuals and organizations who are participating in them, and we are proud to play our part.

[329] Arvind Narayanan, "Translation tutorial: 21 fairness definitions and their politics," tutorial presented at the Conference on Fairness, Accountability, and Transparency, (Feb 23 2018), available at: https://www.youtube.com/embed/jIXIuYdnyyk

6. Bibliography

Access Now, ‘Human Rights in the Age of Artificial Intelligence’ (2018) <https://www.accessnow.org/cms/assets/uploads/2018/11/AI-and-Human-Rights.pdf>
Amnesty International, Access Now, ‘Toronto Declaration: Protecting the Right to Equality and Non-Discrimination in Machine Learning Systems’ (2018) <https://www.accessnow.org/cms/assets/uploads/2018/08/The-Toronto-Declaration_ENG_08-2018.pdf>

Artificial Intelligence Industry Alliance, ‘Artificial Intelligence Industry Code of Conduct (Consultation Version)’ (2019) <https://www.secrss.com/articles/11099>

Beijing Academy of Artificial Intelligence, ‘Beijing AI Principles’ (2019) <https://www.baai.ac.cn/blog/beijing-ai-principles?categoryId=394>

British Embassy in Mexico City, ‘Artificial Intelligence in Mexico (La Inteligencia Artificial En México)’ (2018) <https://docs.wixstatic.com/ugd/7be025_ba24a518a53a4275af4d7ff63b4cf594.pdf>

Chinese National Governance Committee for the New Generation Artificial Intelligence, led by China’s Ministry of Science and Technology, ‘Governance Principles for a New Generation of Artificial Intelligence: Develop Responsible Artificial Intelligence’ (2019) <http://www.chinadaily.com.cn/a/201906/17/WS5d07486ba3103dbf14328ab7.html>

Council of Europe, European Commission for the Efficiency of Justice, ‘European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and Their Environment’ (2018) <https://rm.coe.int/ethical-charter-en-for-publication-4-december-2018/16808f699c>

European Commission, ‘Artificial Intelligence for Europe: Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee, and the Committee of the Regions’ COM (2018) 237 <https://ec.europa.eu/digital-single-market/en/news/communication-artificial-intelligence-europe>

European Commission’s High-Level Expert Group on Artificial Intelligence, ‘Ethics Guidelines for Trustworthy AI’ (2018) <https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai>

Future of Life Institute, ‘Asilomar AI Principles’ (2017) <https://futureoflife.org/ai-principles/?cn-reloaded=1>

G20 Trade Ministers and Digital Economy Ministers, ‘G20 Ministerial Statement on Trade and Digital Economy’ (2019) <https://www.mofa.go.jp/files/000486596.pdf>

German Federal Ministry of Education and Research, the Federal Ministry for Economic Affairs and Energy, and the Federal Ministry of Labour and Social Affairs, ‘Artificial Intelligence Strategy’ (2018) <https://www.ki-strategie-deutschland.de/home.html>

Google, ‘AI at Google: Our Principles’ (2018) <https://www.blog.google/technology/ai/ai-principles/>

IA Latam, ‘Declaración de Principios Éticos Para La IA de Latinoamérica’ (2019) <http://ia-latam.com/etica-ia-latam/>

IBM, ‘IBM Everyday Ethics for AI’ (2019) <https://www.ibm.com/watson/assets/duo/pdf/everydayethics.pdf>

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’ (2019) First Edition <https://ethicsinaction.ieee.org/>

Information Technology Industry Council, ‘AI Policy Principles’ (2017) <https://www.itic.org/resources/AI-Policy-Principles-FullReport2.pdf>

Japanese Cabinet Office, Council for Science, Technology and Innovation, ‘Social Principles of Human-Centric Artificial Intelligence’ (2019) <https://www8.cao.go.jp/cstp/english/humancentricai.pdf>
Microsoft, ‘AI Principles’ (2018) <https://www.microsoft.com/en-us/ai/our-approach-to-ai>

Mission assigned by the French Prime Minister, ‘For a Meaningful Artificial Intelligence: Toward a French and European Strategy’ (2018) <https://www.aiforhumanity.fr/pdfs/MissionVillani_Report_ENG-VF.pdf>
Monetary Authority of Singapore, ‘Principles to Promote Fairness, Ethics, Accountability and Transparency (FEAT) in the Use of Artificial Intelligence and Data Analytics in Singapore’s Financial Sector’ (2019) <http://www.mas.gov.sg/~/media/MAS/News%20and%20Publications/Monographs%20and%20Information%20Papers/FEAT%20Principles%20Final.pdf>

New York Times’ New Work Summit, ‘Seeking Ground Rules for AI’ (March 2019) <https://www.nytimes.com/2019/03/01/business/ethical-ai-recommendations.html>

Niti Aayog, 'National Strategy for Artificial Intelligence: #AI for All (Discussion Paper)' (2018) <https://www.niti.gov.in/writereaddata/files/document_publication/NationalStrategy-for-AI-Discussion-Paper.pdf>
Organisation for Economic Co-operation and Development, 'Recommendation of the Council on Artificial Intelligence' (2019) <https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449>

Partnership on AI, 'Tenets' (2016) <https://www.partnershiponai.org/tenets/>

Smart Dubai, 'Artificial Intelligence Principles and Ethics' (2019) <https://smartdubai.ae/initiatives/ai-principles-ethics>

Standard Administration of China, 'White Paper on Artificial Intelligence Standardization' *excerpts in English published by New America* (January 2018) <https://www.newamerica.org/cybersecurity-initiative/digichina/blog/translation-excerpts-chinas-white-paper-artificial-intelligence-standardization/>

Telefónica, 'AI Principles of Telefónica' (2018) <https://www.telefonica.com/en/web/responsible-business/our-commitments/ai-principles>

Telia Company, 'Guiding Principles on Trusted AI Ethics' (2019) <https://www.teliacompany.com/globalassets/telia-company/documents/about-telia-company/public-policy/2018/guiding-principles-on-trusted-ai-ethics.pdf>

Tencent Institute, 'Six Principles of AI' (2017) <http://www.kejilie.com/iyiou/article/ZRZFn2.html>

The Public Voice Coalition, 'Universal Guidelines for Artificial Intelligence' (2018) <https://thepublicvoice.org/ai-universal-guidelines/>

Think 20, 'Future of Work and Education for the Digital Age' (2018) <https://www.g20-insights.org/wp-content/uploads/2018/07/TF1-1-11-Policy-Briefs_T20ARG_Towards-a-G20-Framework-For-Artificial-Intelligence-in-the-Workplace.pdf>

UK House of Lords, Select Committee on Artificial Intelligence, 'AI in the UK: Ready, Willing and Able?' (2018) Report of Session 2017-19 <https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf>

UNI Global Union, 'Top 10 Principles for Ethical Artificial Intelligence' (2017) <http://www.thefutureworldofwork.org/media/35420/uni_ethical_ai.pdf>

United States Executive Office of the President, National Science and Technology Council Committee on Technology, 'Preparing for the Future of Artificial Intelligence' (2016) <https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf>

University of Montreal, 'Montreal Declaration for a Responsible Development of Artificial Intelligence' (2018) <https://www.montrealdeclaration-responsibleai.com/the-declaration>

cyber.harvard.edu