

Issue Brief

The Inigo Montoya Problem for Trustworthy AI (International Version): Comparing National Guidance Documents

Authors

Emelia S. Probasco

Kathleen T. Curlee

Executive Summary

The United States and four key allies, Australia, Canada, Japan, and the United Kingdom, share common principles for trustworthy AI: accountability, explainability, fairness, privacy, security, and transparency. However, variations in their definitions of these principles, as revealed in their respective policy documents, could substantially affect interoperability and commercial exchanges, and hamper the development of international norms. This brief builds off of our prior work on the use of trustworthy AI terms in the U.S. and scrutinizes their use in the national guidelines and policy statements of these four key American allies.¹ We find:

- High-level principles around responsible AI are similar in spirit, but differ in details. Those detailed differences do not represent substantial disagreements among the countries on trustworthy AI terms but they can become problematic as countries build specific guidance and eventually regulations atop the principles.
- All countries value accountability and aim to hold a human responsible for harm caused by an AI system, but countries vary on who should be accountable. Different expectations about the timeliness of accountability processes or the expectation of compensation will complicate international exchanges.
- For explainability and understandability, countries diverge on two core issues: the audience for the explanation and the expected subject of that explanation. Some countries have specific guidance on which audiences require explanations, while others are broad and vague. AI products in use across these nations will have to account for these varied and country-specific expectations, which may be more inefficient than they are helpful to users. Additionally, conflicts could arise when one nation expects certain data to be included in an explanation, but another country finds the inclusion of that data to be problematic for security or privacy reasons.
- Bias and discrimination are uniform concerns when it comes to the issue of fairness, but otherwise, fairness definitions depend on culture and context. Even among the five allies studied here, there are differences in expectations around the involvement of affected users in defining fairness or pursuing accountability for an unfair system.
- All nations value privacy, but differ on what is considered private, how privacy should be achieved, and who is responsible for protecting privacy. In the case of what should be protected, only the U.S. includes a mention of intellectual

property, for example. With respect to how privacy should be achieved, only the UK guides developers to minimize the retention of private information.

- When it comes to transparency and fairness, all countries are clear on the need to at least disclose to a user that they are interacting with an AI system, if not gain consent from that user before an interaction. However, not all countries agree on the kind of situation that necessitates disclosure or consent.
- Security is often closely linked to other data and cybersecurity policies. However, not all countries include malicious attacks explicitly in their list of concerns.

Our analysis is limited to policy statements as they exist today. Each country would likely consider its governance of AI a work in progress. But this also means there is an opportunity now to influence the development of policies, before they harden into more specific guidance and regulation and while many countries are still interested in creating an international consensus on AI governance.

Table of Contents

Executive Summary	1
Introduction	4
Prior Research.....	4
Motivation	5
Selecting Country AI Guidance Documents.....	6
Key AI Documents Evaluated	7
Terms and Their Varied Definitions.....	11
Accountability	11
Explainability and Understandability	12
Fairness.....	15
Privacy.....	17
Security	18
Transparency.....	18
Other Notable Similarities and Differences.....	21
Recommendations.....	23
Points of Unanimity to Solidify	23
Smaller Gaps that Could be Bridged	24
Coordinate to Clarify Larger Differences	24
Learn from Interesting Differences	25
Conclusion	26
Authors	27
Acknowledgments.....	27
Appendix A: Country Inclusion Criteria.....	28
Appendix B: Text Selection and Analysis.....	29
Endnotes.....	31

Introduction

On the surface, many nations are using common principles to guide the governance of AI. Upon closer examination, however, the common principles are not defined or understood in the same way. This issue brief examines how the U.S., Australia, Canada, Japan, and the United Kingdom compare in their use or explanation of common AI principles in their national AI guidelines. While these countries often use the same or similar words in their key documents, the way those terms are used, explained, or defined varies. Those definitional variations will likely affect international diplomatic and economic exchanges and the emergence of truly global norms.

While laws, treaties, and even international technical standards are frequently the focus of international policymaking, norms and national guidance are critical precedents or stopgaps in the absence of formal agreements. Prior research has also established the impacts of international norms on domestic policy debates, policy adjustments, as well as changes in national institutions.² Focusing on the policy guidance of these key allies gives analysts a window into how international and domestic AI policy may develop as more conversations occur on how to regulate AI.

The United States has made clear its interest in leading a coalition of allies in the development of international norms for governing AI. Washington has demonstrated this interest in numerous high-level AI documents over the past several years, including, but not limited to, the 2022 “Blueprint for an AI Bill of Rights,”³ the “Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy,”⁴ and others as discussed at length below. Within these documents, the U.S. both explicitly and implicitly declares its interest in shaping international norms for AI governance and lays out what those norms should include.

Prior Research

The U.S. is not alone in articulating an interest in shaping global AI governance. One 2019 study found 84 state and non-governmental documents that included statements of principle on AI ethics and governance.⁵ A related study in 2020 analyzed 25 national AI policy documents that made statements on AI ethics.⁶ Each of these policy documents are frequently updated, which complicates an already fast-moving global conversation about AI governance. During the course of research for this paper, for example, three of the national policy documents examined in this study were substantially revised or entirely new guidance was added.

Amidst the proliferation and revision of high-level guidance on trustworthy AI or AI governance, researchers have been looking for trends that signal the emergence of norms. Previous work has identified several common terms. The aforementioned study of 84 documents in 2019, for example, identified five convergent principles for AI development and use: transparency, justice and fairness, non-maleficence, responsibility, and privacy. However, while the words used to title a principle were common, there were substantial variations in the definitions of those principles.⁷ Another 2021 CSET report found agreement on even higher-level concepts, such as complying with existing legal frameworks, human centricity, ethical risk identification, and the need for risk mitigation within defense-oriented policies.⁸ A thorough and more quantitative semantic analysis of the ethical terms in 25 national AI policy documents in 2020 illuminated term uses and documented some definitional differences.⁹ Our paper builds on this previous research and contributes to the broader discussion by investigating the qualitative differences in definitions and the implication of those differences for the implementation of global norms.

Outside of the research community, the U.S. government has also compared its terminology to the terminology of other governments. An early version of the U.S. National Institute of Standards and Technology's AI Risk Management Framework (NIST AI RMF)¹⁰ included a mapping of the key characteristics listed in that document to the terms used in other high-level U.S. guidance documents as well as the Organisation for Economic Co-operation and Development (OECD) AI Principles¹¹ and the proposed European Union Artificial Intelligence Act (EU AI Act).¹² Unfortunately, the mapping of terms was missing from the final version of the NIST AI RMF, which is understandable given the challenge of keeping up with policy document changes and the variations in definitions.

Motivation

Despite the difficulty, however, these comparisons of AI policy frameworks and definitions are useful to policymakers who are working to establish partnerships and alliances that will advocate for international AI governance norms. This paper attempts to address this need by providing policymakers with specific insights on the substance and relevance of differences in definitions of key trustworthy AI terms. How countries define AI ethics can have a significant impact on how the AI works, as well as the situations in which it is deployed. For example, in 2016, there was a large public dispute between the news organization ProPublica and the algorithm company Northpointe over the fairness and alleged bias of an algorithm used to support bail and sentencing decisions. To some analysts, the debate had less to do with the statistical approach than with what was viewed as fair: in other words, is it fair if the algorithm

predicts outcomes equitably across races or is it fair if the algorithm makes errors equitably across races? Mathematically, both cannot be simultaneously satisfied.¹³ It is for this and similar reasons that we seek to address definitional differences by countries in trustworthy and ethical AI keywords before the differences become ingrained and problematic for users, companies, and nations alike.

Selecting Country AI Guidance Documents

To scope our research, we chose to limit our analysis of terms and definitions to the U.S. and four key allies: Australia, Canada, Japan, and the UK. Each of these countries adhere to the OECD Recommendation of the Council on Artificial Intelligence, which includes a list of “principles for responsible stewardship of trustworthy AI.”¹⁴ Often when creating their guidelines, countries have referred to the EU and OECD’s ethical AI principles as the foundation for their own. Some nations even adopt the OECD guidance and do not publish their own policies at all.¹⁵ For Australia, Canada, Japan, the UK, and the U.S., however, each has published specific national guidance on the governance of AI.

For the countries analyzed, there are numerous AI strategy, governance, or policy documents written by different agencies across each government. In general, this analysis prioritizes documents that appear most likely to influence future AI policies. In reviewing these documents, official policies issued by the highest executive agency available were prioritized. In some cases, voluntary frameworks for countries that have yet to articulate a binding national policy were included. The analysis, therefore, compares documents that include a mix of voluntary frameworks, government agency policies, government guidance on laws, and laws that have different objectives and purposes. The comparison is admittedly imperfect, but the comparison of definitions within those documents is still highly relevant to understanding how countries will approach AI governance in future laws and policies as they are developed. The documents selected for analysis for each country are listed below, along with very brief descriptions of those documents. A more detailed description of our method of selecting these five countries can be found in Appendix A. For a more complete explanation of our approach to analyzing the text of each document, please see Appendix B.

Key AI Documents Evaluated

Australia

Australia's AI Ethics Principles

“Australia’s AI Ethics Principles,” published by the Department of Industry, Science, and Resources is a voluntary framework designed to complement relevant laws and regulations and guide both public and private uses of AI.¹⁶ The document provides a brief overview of eight principles that contribute to ethical AI as well as detailed descriptions of each. The document’s goal is to ensure positive outcomes of AI uses while also building public trust. Of note, the document cites the Institute of Electrical and Electronics Engineers (IEEE) report, “Ethically Aligned Design,” as a source of inspiration and guidance.¹⁷

Canada

Directive on Automated Decision-Making

Canada’s “Directive on Automated Decision-Making,” is part of its broader “Policy on Service and Digital,” which was released in March 2019 and updated in 2021 and 2023. It is intended to regulate the Government of Canada’s use of any AI-enabled “system, tool, or statistical models used to recommend or make an administrative decision about a client.”¹⁸ The Directive builds upon prior Canadian legislation, which includes the “Policy on Government and Digital” and the “Privacy Act.” Notably, the Directive includes impact assessment levels that guide the application of governance requirements to AI. Briefly, these levels are:

- Level I: “Decisions will often lead to impacts that are reversible and brief.”
- Level II: “Decisions will often lead to impacts that are likely reversible and short-term.”
- Level III: “Decisions will often lead to impacts that can be difficult to reverse and are ongoing.”
- Level IV: “Decisions will often lead to impacts that are irreversible and perpetual.”¹⁹

Responsible Use of Artificial Intelligence-Guiding Principles

The “Directive on Automated Decision-Making” is augmented by Canada’s “Responsible Use of Artificial Intelligence-Guiding Principles.”²⁰ The AI Guiding Principles are aligned with Canada’s administrative law principles and were developed in conjunction with the “Directive on Automated Decision-Making” through a series of workshops and whitepapers. These sessions included numerous Canadian government agencies as well as input from industry and academia.

Japan

Social Principles of Human-Centric AI

“The Social Principles of Human-Centric AI” were created as a part of Japan’s Society 5.0 in 2019, which sets as a goal the creation of “a sustainable human-centric society that implements AI, IoT (Internet of Things), robotics, and other cutting-edge technologies to create unprecedented value, and a wide range of people can realize their well-being while respecting the well-being of others.”²¹ The Social Principles are voluntary but are designed to impact AI guidelines for Japanese companies.

Governance Guidelines for Implementation of AI Principles

The Social Principles also influence the “Governance Guidelines for Implementation of AI Principles,” which are an amalgamation of domestic and international trustworthy AI guidelines and recommendations.²² Released on January 28, 2022, by the “Expert Group on How AI Principles Should be Implemented,” the Guidelines are designed to be the main reference point for Japanese companies when developing governance mechanisms for AI. The document is composed of action targets and implementation examples with a particular focus on AI systems that could negatively impact society.²³

While the principles listed are not legally binding, the Guidelines are intended to supplement the voluntary efforts of companies to develop in-house trustworthy AI guidelines. Similar to corporations in other countries, in the absence of legally binding regulations on trustworthy AI, Japanese companies are expected to ensure their products adhere to other laws not focused on AI, such as non-discrimination laws.

The United Kingdom

[Information Commissioner's Office \(ICO\) "Guide to the UK General Data Protection Regulation and Data Protection Act"](#)

While the UK General Data Protection Regulation and Data Protection Act (UK GDPR) is not originally targeted at artificial intelligence, many of its principles apply to AI by virtue of the technology's reliance on data.²⁴ To better explain the requirements embedded within UK GDPR to government practitioners, the Information Commissioner's Office (ICO) produced the "Guide to the UK General Data Protection Regulation" with much more specific guidance on how the UK GDPR applies to AI systems.²⁵

The ICO Guide explains key principles in the UK GDPR and includes checklists for each principle, which give greater insights into the intended meaning of each keyword. While we examined both the GDPR and its implementation guidance, we selected the "Guide to the UK GDPR" instead of the UK GDPR itself, as the former consolidates and breaks down each relevant principle in a way that is digestible to practitioners who work with AI daily.

[ICO and The Alan Turing Institute "Explaining Decisions Made with AI" and ICO "Guide on AI and Data Protection"](#)

The "Guide to the GDPR" is also updated regularly to accommodate any relevant legislative or technical changes that may affect AI systems or their governance. Those updates include substantial additions to the guidance on "Explaining Decisions Made with AI" in October 2022,²⁶ which was a collaboration between the ICO and The Alan Turing Institute, and the "Guidance on AI and Data Protection," which was updated in March 2023.²⁷ The drawback of including the ICO guidance on the UK GDPR is that the guidance document is far more detailed than many of the other policies or high-level documents included in this study. We accepted this difference because the guidance is drawn directly from the UK GDPR and explicitly links to high-level principles, which makes it one of several UK government documents that illustrate the influence of high-level principles on more detailed guidance.²⁸

The United States

Presidential Executive Order 13960

The President of the United States issued Executive Order 13960, “Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government” on December 3, 2020.²⁹ Executive Orders in the U.S. manage federal operations and direct federal entities to take specific actions. They are enforceable and have the effect of law. EO13960 specifically directed federal agencies to “design, develop, acquire, and use AI in a manner that fosters public trust and confidence while protecting privacy, civil rights, civil liberties, and American values, consistent with applicable law and the goals of Executive Order 13859.”³⁰

White House Office of Management and Budget “Guidance for Regulation of Artificial Intelligence Applications”

Published in 2019, Executive Order 13859, “Maintaining American Leadership in Artificial Intelligence,” provides guidance to executive agencies on how to support the research and development of AI-enabled systems, and only mentions ethical AI keyword terms in passing.³¹ The Office of Management and Budget, however, published amplifying guidance on EO13859, titled “The Guidance for Regulation of Artificial Intelligence Applications,” which specifically addresses key terms. Accordingly, the OMB guidance, and not EO13859, was included in our analysis, alongside EO13960.³²

Alongside EO13960 and the OMB “Guidance for Regulation of Artificial Intelligence Applications,” there are many other documents on responsible AI issued by the U.S. government. While a full inventory of executive branch documents about responsible and trustworthy AI is beyond the scope of this paper, one bears special mention: the White House Office of Science and Technology Policy’s “Blueprint for an AI Bill of Rights,” released in October 2022.³³ This document contains definitions of trustworthy AI and attempts to draw a more explicit connection between the principles of trustworthy AI and core democratic values. The document is non-binding, however, and does not reflect the U.S.’ position on AI ethics, nor should it be used to represent the position the U.S. will take in international collaborations. Accordingly, we reviewed the document but ultimately chose to exclude it from this analysis.

Terms and Their Varied Definitions

Six terms appeared consistently in high-level governmental guidance documents we examined: accountability, explainability, fairness, privacy, security, and transparency. This finding reinforces previous research which has found the common terms to be: responsibility, transparency, justice and fairness, privacy, and non-maleficence.³⁴

Having identified the six common terms, the following analysis focuses on how they were used in each country's key documents.³⁵ It is important to note that where differences are highlighted, they do not represent substantial disagreements among the countries on trustworthy AI terms. In fact, there appears to be broad agreement, which gives reason to believe that nuanced differences could be overcome in an effort to create more global policies.

Accountability

One consistent theme when the five nations discussed accountability was that humans must be accountable for the adverse outcomes of AI systems for which they bear some responsibility. Where nations vary is on the importance and the role of a human operator, the role of an affected person in an accountability process, and the specific designation of accountability within the government when a government agency uses AI. Among the differences noted in the use of the term accountability:

Human Intervention

Australia, Canada, and Japan all indicate a need for human intervention in the operation and deployment of an AI system in the event that an AI system causes harm. In the case of Australia, there is an expectation that “human oversight of AI systems





Human Intervention				
				

should be enabled” and that organizations must “consider the appropriate level of human control or oversight for the particular AI system or use case.”³⁶ Canada’s Directive is more specific and includes guidance that humans should be able to intervene in level III and level IV AI systems both in advance of system deployment and during operations.³⁷ Therefore, both nations

seem to indicate that an operator capable of stopping an AI system that is actively harming users is accountable for doing so. Japan’s guidance calls for allocating “responsibilities to those who are able to mitigate negative impacts.”³⁸ This could be viewed as similar to the Australian and Canadian guidance, but mitigation could occur before, during, or after an incident, and not just by stopping the AI system entirely. All of these requirements, however, are somewhat vague and sidestep the still ongoing

debate about the proper role of humans in AI system operations. In other words, should humans be “in the loop,” approving and rejecting all actions, or should they be “on the loop” observing the AI system in action and only intervening when required?

Role of the Affected Person

Affected Person Role				
				

Australia,³⁹ Japan,⁴⁰ and the UK,⁴¹ all note the person affected by the AI must be part of any accountability processes.⁴² The centrality of the affected person in an accountability process is echoed in each country’s conception of fairness (see the section on fairness). Australia⁴³ and the UK⁴⁴ specifically emphasize the need for affected persons to be able to challenge an AI’s decisions. The UK ICO goes further and says processes and results must be documented to an “auditable standard” for accountability.⁴⁵ Additionally, Australia includes the potential for compensation and a timely accountability process for those harmed by an AI.⁴⁶

Government Accountability

Government Accountability				
				

While all five countries highlight the importance of accountability, only two countries delineate a process for assigning it when government agencies use AI. In Canada’s “Directive on Automated Decision-Making,” responsibility for fulfilling the responsible AI requirements within the Canadian government is assigned to the Assistant Deputy Minister responsible for the program that will use the automated system or their named designee.⁴⁷ In the UK’s “Guidance on AI and Data Protection,” data protection officers are called out as being directly responsible for data risk management and governance of AI systems.⁴⁸ Data protection officers are also accountable for understanding the GDPR and its impact on AI tools and systems.

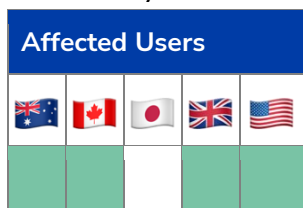
Explainability and Understandability

All countries in our study discuss either explainability, understandability, or both, often in the context of other key principles such as transparency, accountability, interpretability, and fairness.⁴⁹ Each country varies in its expectations around the concepts and the UK’s “Guidance on explaining AI” stands out as the most detailed and far-reaching. Overall, the main issues for explainability center on questions of who receives the explanation and what should be explained.

Specific guidance on the audiences for AI explanations can guide developers to create approaches designed for those audiences and their circumstances. Crafting appropriate explanations for a given audience is non-trivial work and establishing common expectations about the types of audiences (i.e., users versus auditors) may better support exchanges of information between allies and cross-border interoperability. That said, country expectations for audiences vary significantly today. Japan is the most limited in its audience expectations, stating simply that explanations should be provided on a “case by case basis.”⁵⁰ This stands in contrast to Australia, for example, which includes users, creators, legal representatives, and the public in its list of audiences affected.⁵¹ The UK, and the U.S. also include other audiences in addition to users. The UK’s guidance specifies that staff whose decisions are supported by an AI system are entitled to a sufficient explanation, as are auditors or external reviewers.⁵² The U.S. issues a blanket statement for explanations to “others, as appropriate.”⁵³ The large blanket statements (i.e. Australia’s “the public” or America’s “others as appropriate”) may help future-proof policies from changing norms and technical abilities, but the ambiguity also creates challenges because audiences have different levels of understanding of AI systems. Moreover, there could be frequent requests for explanations from diverse parties with multiple and varying motivations.

Affected Users

All but Japan expect that affected users will be provided an explanation of the decision of an AI system. Notably, under the GDPR, the UK explicitly encourages developers




and operators of AI systems to consider children or other vulnerable groups in preparing explanations for affected users.⁵⁴

The UK recommendation to include explanations accessible to vulnerable groups is unique within the documents we reviewed but the sentiment aligns with common notions of equal opportunity and anti-discrimination across all five countries.






Method of Explanation Delivery

The UK is unique in including explicit guidance on who should deliver the explanation of an AI system, stating that the information should be delivered as a conversation and that “people should be able to discuss a decision with a competent human being.”⁵⁵

Method of Explanation Delivery				
				

What to Explain: Notification, System Structure, and System Outcomes

Countries generally recognize three points that require explanation: first, an explanation that an individual is interacting with an AI-based system and the role of that AI system in a decision (related to the notion of notification or informed consent as explained in the section that follows on transparency). Second, an explanation of how the system works. And third, an explanation of the system’s output or decision.

What to Explain				
				

The UK and Canada embrace all three points for explanation: notification, system structure, and system outcomes. For example, the UK guidance states that individuals have the right to be informed that they are interacting with an automated system for decision-making; provided information about the logic involved in the system and how the system may impact the individual; and, after a decision is made, given an explanation of the result.⁵⁶ Canada adds to this list a requirement to explain the training data for the system and, if applicable, the way it was collected.⁵⁷ Both Canada and the UK further delineate expectations for explainability based on the impact level of an AI system. In Canada, for example, AI systems that have reversible and brief impacts have a lower expectation for explainability than systems that have irreversible or perpetual impacts.⁵⁸ The other countries are less precise about these requirements:

- The U.S. applies a standard of understanding to both the operations of the system and its outcomes.⁵⁹
- Australia⁶⁰ and Japan⁶¹ mention a need for the explainability of system outcomes or results, but not necessarily how the system works.

Specific Guidance on Explainable Approaches

Specific Guidance on Explainable Approaches				
				






Unlike other nations, the UK's guidance on explaining AI goes into great detail on different types of explanations (i.e., rational, responsibility, data explanations, etc.) as well as types of AI models that lend themselves to better explanations (i.e., a linear regression model vs. an artificial neural net).⁶² This level of detail is unique among the documents reviewed. The UK's guidance on explainability could serve as a useful guide for companies or government agencies deploying AI.

Fairness





Fairness was a consistently emphasized theme in the AI guidance documents we examined. Although the term is common, the definitions predictably vary, since concepts of fairness also vary by geographic and cultural norms.⁶³ Additionally, fairness is a difficult principle to define mathematically, morally, or politically. While fully agreeing on common standards for fairness may be a bridge too far, there are similarities worth noting among the five countries, especially the importance of engaging an affected user and preventing discrimination.

Role of Affected User

Australia,⁶⁴ Japan,⁶⁵ the UK,⁶⁶ and the U.S.⁶⁷ include affected users as being parties to defining and judging the fairness of an AI system. "Affected users" include (in the case of all four listed countries) the individuals who may be affected by the decisions of an AI system, the individuals who may interact with an AI system, as well as the individuals whose data may have been used to train or maintain the system. While Canada does mention affected users in meeting explainability requirements, it does not do so in terms of fairness.

Role of Affected User				
				

Importance of Disclosure or Consent

Importance of Disclosure or Consent				
				

Australia⁶⁸ and the UK⁶⁹ are especially clear on the need to elicit informed consent from users who may interact with an AI system. This is echoed in other national discussions of transparency (see section on transparency). Aside from the requirement for informed consent in advance of an interaction, however, no country has yet defined the specific method by which affected users will be notified or engaged in a process.

Bias and Discrimination






Discrimination features in the definitions of fairness for Australia,⁷⁰ Japan, the UK,⁷¹ and the U.S.⁷² Only Japan goes so far as to mention specific categories that should be protected such as age, gender, nationality, race, and religion.⁷³ The other countries do not list specific protected classes but instead emphasize inclusiveness and accessibility.

Bias and Discrimination				
				

Of note, Japan⁷⁴ and the U.S.⁷⁵ emphasize the democratic notions of civil rights or civil liberties in their definitions of fairness. Australia, Canada, and the UK do mention civil rights or liberties in other portions of their documents but not in relation to fairness. This inclusion of democratic norms is discussed more below.

Procedural Fairness

Canada is somewhat unique in that it does not discuss discrimination, but instead draws upon its established concept of procedural fairness. In

Procedural Fairness				
				

Canada, any applicant for government resources or a government decision is entitled to a decision “free from a reasonable apprehension of bias, by an impartial decision-maker.”⁷⁶ The procedural standard also includes, among others, expectations that decisions will be processed without undue






delay, that the applicant has a right to be heard in response to a decision, and that the applicant has a right to be told the reasons for the decision.⁷⁷ This notion of procedural fairness informs Canada’s approach to transparency and explainability as well.

Privacy

While all five countries include the term “privacy” in their top documents and many references to established guidelines on data protection, there are a few interesting differences in their statements, mostly about what is to be protected (should it include intellectual property?), how (by security or by a data minimization standard?), and why (should privacy be characterized as a democratic value?).

Intellectual Property






The U.S. is the only country to make specific mention of intellectual property in conjunction with privacy.⁷⁸ This may be linked to the U.S. assumption that safeguarding intellectual property is foundational to economic growth. It is also possibly linked to

Intellectual Property				
				

America’s vocal concerns about the theft of intellectual property by China, declared by former FBI Director Christopher Wray as the nation’s “greatest long-term threat.”⁷⁹ American allies also recognize the importance of intellectual property and have made similar declarations regarding China’s theft of IP, but do not mention it explicitly in their descriptions of privacy and AI.⁸⁰






Data Minimization

The UK highlights data minimization as a method for enhancing privacy. It notes that

Data Minimization				
				

“personal data shall be adequate, relevant, and limited to what is necessary in relation to the purposes for which they are processed.” The UK also includes guidance to conduct due diligence on any third-party services to ensure that privacy is maintained when relying on a vendor for either data or AI systems.⁸¹

Privacy and Democratic Values

Privacy and Democratic Values				
				






Japan,⁸² the UK,⁸³ and the U.S.⁸⁴ all link privacy to individual rights and freedoms. However, the U.S. mentions the link most often in their AI documents. In EO13960, the phrase “privacy, civil rights, civil liberties” is used five times, and in two instances the phrase “American values” is included. Japan states “we should make sure that any AI using personal data and any service solutions that use AI, including use by the government, do not infringe on a person's individual freedom, dignity or equality.”⁸⁵ Other countries also mention

democratic values, but not as an aspect of privacy. For example, Australia highlights democratic values as a component of its “human-centered values” principle.⁸⁶

Security






Although all five countries mention security frequently across their AI guidance documents, the term is generally referenced as a component of other keywords rather than as an independent principle. This may be because security is often addressed in relation to cyber or data-security policies and requirements. Still, all five countries uniformly accept the need for a risk management approach to security, and three of the five countries include guidance to build and operate systems in a way that fortifies them against an attack.

Risk Management Approach

Risk Management Approach				
				

All five countries explicitly address security concerns through risk assessment or risk management frameworks and processes. The documents make clear that AI systems contain risk and that governance is a process of *managing* risk, as opposed to eliminating it.

Preparing for an Attack or Breach

Risk Management Approach				
				

Australia,⁸⁷ the UK, and the U.S.⁸⁸ all show concern for malicious attacks against AI systems. Australia and the U.S. specifically mention the requirement for resilience, which is to say that AI systems should have various backup options or what is termed graceful degradation in the event of an attack. The UK states there must be “appropriate levels of security against [data’s] unauthorized or unlawful processing, accidental loss, destruction or damage.”⁸⁹





Transparency

Transparency is related to, but distinct from, explainability. The NIST AI RMF 1.0 definition is helpful in providing clarity here, it defines transparency as “the extent to which information about an AI system and its outputs is available to individuals interacting with such a system” and explainability as “a representation of the mechanisms underlying AI systems’ operation.”⁹⁰ Not all countries share this clear distinction in their high-level policy documents, and the lack of clarity can be confusing.

There are two components to transparency as distinct from explainability that appear in the examined policy documents: one that has to do with unanimous support for disclosure for eliciting user consent which, to an extent, overlaps with explainability, and another relates to the ability to observe the workings of the AI system.

Disclosure

Australia,⁹¹ Canada,⁹² Japan,⁹³ the UK,⁹⁴ and the U.S.⁹⁵ all emphasize the importance of providing notice to a user that they are interacting with a system that uses AI to make decisions. The timing and method of disclosure are vague, but the U.S. does include guidance that “...disclosures, when required, should be written in a format that is easy for the public to understand.”⁹⁶ This requirement is related to the previously discussed principles of explainability and fairness, though in this instance all countries agree on the need to disclose as a part of the principle of transparency.

Disclosure				
				






Canada's approach to providing notice is different because it does not require disclosure for systems that only have reversible and brief impacts (Level I). Higher-level systems, whose impacts can range from reversible and short-term to irreversible and perpetual, require disclosure.⁹⁷

Balancing Transparency with Privacy






Balancing Transparency with Privacy				
				

The U.S.⁹⁸ and Canada⁹⁹ recognize an inherent tension between transparency and two other principles they value: security and privacy. Canada states this tension well, saying that the government will “be as open as we can by sharing source code, training data, and other relevant information, all while protecting personal information, system integration, and national security and defence.”¹⁰⁰

Table 1. Summary Table of AI Themes and Principles Contained in Guidance Documents, by Country

Theme	Specific Principle					
Accountability	Human Intervention	■	■	■		
	Role of Affected Person			■	■	
	Government Accountability		■		■	
Explainability and Understandability	Affected Users (Who)	■	■		■	■
	Method of Explanation Delivery (What)				■	
	What to Explain: Notification, System Structure and Outcomes		■		■	
	Specific Guidance on Explainable Approaches				■	
Fairness	Role of Affected User	■		■	■	■
	Importance of Disclosure or Consent	■			■	
	Bias and Discrimination	■		■	■	■
	Procedural Fairness		■			
Privacy	Intellectual Property					■
	Data Minimization				■	
	Privacy and Democratic Values			■	■	■
Security	Risk Management Approach	■	■	■	■	■
	Preparing for an Attack or Breach	■			■	■
Transparency	Disclosure	■	■	■	■	■
	Balancing Transparency with Privacy		■			■

Other Notable Similarities and Differences

Additional Terms					
					
Accuracy					
Reliability					
Safety					
Robustness					
Democratic Values					

While this issue brief’s goal is to analyze the similarities and differences among the five countries’ shared ethical AI terms, our research revealed a number of terms that several, but not all, countries shared.

Accuracy

The academic¹⁰¹ and technical¹⁰² communities have written extensively about the importance of accurate and reliable AI, but unfortunately, not every country lists accuracy as a core principle. There are too¹⁰³ many¹⁰⁴ examples¹⁰⁵ of deployed systems with substantial accuracy issues that, at least in some cases, render their predictions as effective as tossing a coin. Australia, the UK, and the U.S. include accuracy in their lists of principles. The UK further distinguishes statistical accuracy for an AI system from the GDPR’s “accuracy principle,” which is related but distinct because it requires that data held or used by the government to be accurate. Japan mentions accuracy as a metric for assessment (both prior to deployment and as a part of a maintenance cycle) but not as a principle. Canada only briefly notes that systems should be accurate, as should the data used to train the system.¹⁰⁶

The absence of accuracy as an independent principle for Japan and Canada may have several explanations, and among them could be the presumption that accuracy is so obvious it need not be stated, or that an inaccurate system is, in fact, an unfair system as in the case of the UK. Given the still frequently reported instances of accuracy errors and their harms, however, countries may do well to further elevate their concerns about accuracy and articulate reasonable expectations for AI accuracy. That said, perfect accuracy is a lofty if not impossible goal and one human decision-makers routinely fail to achieve. Overall, more specificity regarding expectations of accuracy may warrant further discussion.

Reliability

Related to accuracy, Australia,¹⁰⁷ Japan,¹⁰⁸ the U.S.,¹⁰⁹ and the UK highlight reliability in ways that align with the International Standards Organization's definition as the "ability of an item to perform as required, without failure, for a given time interval, under given conditions."¹¹⁰ However, the UK adds a point about the need to ensure the data comes from a reliable source, in addition to the reliability of an AI-enabled system.¹¹¹

Safety

Only the U.S. and Australia include safety as a unique keyword term for ethical or trustworthy AI. In the U.S., safety is linked to security and couched in terms of eliminating security vulnerabilities or the potential for malicious use.¹¹² In Australia, safety is linked to reliability and the idea that systems should operate "in accordance with their intended purpose."¹¹³ While Canada does not refer to safety as a key principle, it does identify the need to "safeguard against unintentional outcomes" as a matter of quality assurance.¹¹⁴

Robustness

Robustness is not addressed more than in passing in any of the high-level policy documents we analyzed. The International Organization for Standardization/International Electrotechnical Commission (ISO/IEC) defines robustness as the "ability of a system to maintain its level of performance under a variety of circumstances."¹¹⁵ Like ISO/IEC, NIST highlights the importance of robustness in the AI RMF 1.0 and includes it as a concept related to accuracy and reliability. The absence of robustness from policy documents may be because the term is thought of as more technical, and less of a stand-alone principle. Robustness is indeed connected to accuracy and reliability, but given the broad technical concerns about the brittleness of AI systems, policymakers may wish to consider the value of elevating robustness as a term within non-technical communities and documents.¹¹⁶

Democratic Values

While no country had a singular ethical AI principle named "democratic values," many of the country documents analyzed contained the phrases "individual rights," or "individual freedoms," and a few included "civil liberties" and/or "democratic values." The desire for democratic nations to promote AI technologies that abide by democratic principles is unsurprising, though there may be some disagreement about how prominently the words should feature in documents meant to influence international

norms. Among the countries analyzed, Canada,¹¹⁷ Japan,¹¹⁸ the UK,¹¹⁹ and the U.S.¹²⁰ are well aligned and emphasize civil rights or civil liberties in their definitions of other terms, particularly fairness. Australia words its support differently, stating “human rights, diversity, and the autonomy of individuals” should be respected by AI systems, though the notion is still consistent with the other nations.¹²¹

Recommendations

To help shape global norms for governing AI that will ultimately affect international commerce, diplomacy, and interoperability, the U.S. will need to monitor the statements of like-minded nations and hone the current common sentiments into specific multilateral agreements and guidelines. That work can start by solidifying current points of unanimity and by seeking out opportunities to bridge narrow gaps. At the same time, the U.S. should deeply engage in areas of more substantial differences and should learn from the efforts of allies who are taking complementary approaches:

Points of Unanimity to Solidify

While they have different reasons, each nation in this study shares statements about the importance of engaging affected users and disclosing and/or seeking consent before the user engages an AI system that could affect them. **The U.S. and its allies should build on this consensus and define how and when notifications should take place and the degree of impact (or the kind of AI) that would trigger the need for a notification.**

The unanimity around the need for security and the use of risk management approaches can also represent an important, if imperfect, opportunity. While nations may not agree on the method of risk management, **instantiating a norm whereby all countries have a publicized and transparent risk management approach seems prudent and possible.**

Finally, there is unanimous agreement that AI systems should support democratic values and individual rights and freedoms. The U.S. and Japan are the most outspoken in the use of the terms but all countries make their preferences clear. Similarly, they also all support a principle that AI systems should avoid causing or perpetuating unfair discrimination. Explicitly listing democratic values in a global list of AI principles may be a bridge too far, given geopolitical differences. **However, these five nations should work together to solidify and explicitly acknowledge the centrality of democratic values in guiding their national AI policies.**

Smaller Gaps that Could be Bridged

Most countries believe in human accountability for AI harms, but **all countries should choose to make explicit the role and method of engaging an affected user in an AI accountability process.**

The details of whether a human is “in the loop” or “on the loop” is one that likely requires more debate and will be differentiated by the perceived risk of the AI system and its application. **However, given current national statements about accountability principles, all nations should choose to embrace the notion of holding a person accountable who is in a position to stop an active AI system from perpetuating harm and ensure the technology is developed with an oversight capability that allows a human supervisor to rapidly intervene.**

Relatedly, for users to meaningfully engage in an accountability process, they need AI to be explainable and understandable. The countries examined in this analysis are nearly unanimous on the need for the explainability of an outcome or decision to users affected by that decision. And even as the outlier, Japan’s stance to provide explanations on a “case-by-case basis” does not seem opposed to those of the other nations examined. Therefore, **all nations should consider agreeing that explainability is geared toward, at a minimum, affected users. Including vulnerable populations, such as children and the elderly, in the affected user category also seems achievable given the statements of all five nations.**

Beyond explaining outcomes, the agreement of Canada, the UK, and the U.S. on the need to explain how the system works (in addition to its outcome) provides a basis for further developing internationally shared notions of explainability.

Coordinate to Clarify Larger Differences

A tension exists between the principles and expectations of transparency and privacy for AI systems. The balance between the two principles will be challenging for governments, citizens, and AI developers alike. **Leaders should engage broadly to collaboratively evolve expectations about the boundaries and balance between transparency and privacy, which will be important to the acceptance and trust of AI systems writ large.**

Learn from Interesting Differences

U.S. policymakers should consider the utility of the UK’s approach to disseminating specific but voluntary guidance on ways developers may comply with the spirit of core AI principles. If this approach can help enable trust in AI systems and improve compliance with high-level principles, without stifling innovation, it could be a model well worth imitating. In particular, the UK’s “Explaining Decisions Made with AI” can help developers or companies who lack this specific area of expertise comply with the principles of explainability and transparency while still giving room for the creation of new explainability techniques.

Currently, U.S. regulations do not differentiate AI systems by the type of impact they might have, whereas Canada and the UK do. This differentiation by impact may be advantageous to the U.S. and others with a similar approach given how challenging it will be to meet the technical and/or administrative burden imposed by some of the principles reviewed in this analysis. **The U.S. and other nations may wish to consider the advantage of adopting a more explicitly risk-based approach to common, core principles.** Differentiating AI systems by risk level could also help the international community focus its efforts on developing norms for those AI systems most concerning to governments.¹²²

Conclusion

Through careful analysis of American, Australian, British, Canadian, and Japanese high-level policies and guidance on trustworthy artificial intelligence, one can find that the U.S. and these key allies generally agree on the importance of six concepts—accountability, explainability, fairness, privacy, security, and transparency— but they do not align on the specific features and definitions of each term. As those terms and definitions evolve, differences will be accentuated, ultimately making commerce, interoperability or international use of AI systems difficult. Without common expectations around notifications, explanations, or accountability, companies and individuals may find it difficult if not impossible to efficiently develop AI solutions that international audiences will trust or that will comply with a myriad of different international policies.

The terms and explanations extracted from each country's documents are still evolving, but as time passes these principles will shape more specific policies and standards across a wide variety of applications. The time to shape the development of these policies is now, while the differences in the high-level principles still do not represent dramatic departures and while allies are still willing to learn from and evolve with each other towards international norms for the governance of AI.

Authors

Emelia Probasco is a Senior Fellow at the Center for Security and Emerging Technology where Kathleen Curlee is a Research Analyst.

Acknowledgments

For feedback and assistance, we would like to thank Catherine Aiken, Alexander Babuta, Tessa Baker, Adam Bartley, Dale Brauner, Samuel Bresnick, Shelton Fitch, Wyatt Hoffman, Margarita Konaev, Jason Ly, and Helen Toner.



© 2023 by the Center for Security and Emerging Technology. This work is licensed under a Creative Commons Attribution-Non Commercial 4.0 International License.

To view a copy of this license, visit <https://creativecommons.org/licenses/by-nc/4.0/>.

CSET Product ID #: 20220039
Document Identifier: doi: 10.51593/20220039
Document Last Modified: 11 October 2023

Appendix A: Country Inclusion Criteria

We selected five countries for comparative analysis based on:

- Membership in either the North Atlantic Treaty Organization, the Quadrilateral Security Dialogue, or the UK – United States of America Agreement (otherwise known as the Five Eyes).
- The presence of high-level policy documents on artificial intelligence that include ethical terminology with defined terms. These documents were largely found through the OECD Artificial Intelligence Policy Observatory.¹²³
- Significant public and/or private sector investments in artificial intelligence, as indicated by the top 15 investing countries in the Emerging Technology Observatory’s Country Activity Tracker.¹²⁴ In most cases, the selected countries are the top AI producers of their region based on investments and research contributions by both the public and private sectors.

These criteria identified nine nations (Australia, Canada, France, Germany, Japan, the Netherlands, the United Kingdom, the United States, and the Republic of Korea) of which the final five were selected for geographic diversity.

Appendix B: Text Selection and Analysis

For each country analyzed, high-level documents included for examination were vetted as being those that:

- were published by an institution charged with the highest level of guidance and governance within that nation. For example, an executive order of the President of the United States;
- provided guidance regarding the development and/or use of AI by the government or by private entities; and,
- included specific references to ethical or trustworthy AI terms and principles in a substantial way.

For each document studied, the authors analyzed the text to identify common terms and explanations related to AI trustworthiness. These terms and explanations were identified, labeled, and grouped through successive reviews, initially using the online text analysis platform Dedoose. For an example of the coding enabled by the Dedoose platform, see Figure 1.

Figure 1: Sample Text Analysis for Fairness

United States: When considering regulations or non-regulatory approaches related to AI applications, agencies should consider, in accordance with **law**, issues of **fairness** and **nondiscrimination** with respect to **outcomes** and decisions produced by the AI application at issue, as well as whether the AI application at issue may reduce levels of **unlawful**, **unfair**, or otherwise **unintended discrimination** as compared to existing processes.

United Kingdom: Second, if you use an AI system to infer data about **people**, in order for this processing to be **fair**, you need to ensure that the system is sufficiently **statistically accurate** and **avoids discrimination**; and you need to consider the impact of **individuals'** reasonable expectations.

Canada: **Decisions** made by **federal government departments** are data-driven, responsible, and comply with **procedural fairness** and due process requirements...**Procedural fairness** is a guiding principle of **governmental and quasi-judicial decision-making**. The degree of **procedural fairness** that the law requires for any given **decision-making** process increases or decreases with the significance of that **decision** and its impact on **rights** and interests.

Australia: Throughout their **lifecycle**, AI systems should be **inclusive** and **accessible**, and should not involve or result in unfair **discrimination against individuals, communities or groups**. This principle aims to ensure that AI systems are **fair** and that they enable **inclusion** throughout their entire **lifecycle**. AI systems should be **user-centric** and **designed** in a way that allows **all people interacting with it to access the related products or services**. This includes both appropriate consultation with **stakeholders**, who may be affected by the AI system throughout its **lifecycle**, and ensuring **people** receive **equitable access and treatment**.

Japan: Under **AI's design** concept, 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.

Color Coding Key:

Law

Fairness

Discrimination

Users

Statistics

People/Individuals

Inclusivity

Procedural Fairness

AI Lifecycle

Accessibility

AI System Design

Outcomes

Unjust/Unlawful

Human Rights

Endnotes

¹ Emelia Probasco, Autumn Toney, and Kathleen Curlee, "The Inigo Montoya Problem for Trustworthy AI" (Center for Security and Emerging Technology, June 2023). <https://doi.org/10.51593/20230014a>

² Andrew P. Cortell and James W. Davis, Jr., "Understanding the Domestic Impact of International Norms: A Research Agenda," *International Studies Review*, 2, no. 1 (Spring, 2000): 65-87, <https://www.jstor.org/stable/3186439>.

³ Office of Scientific and Technology Policy, "Blueprint for an AI Bill of Rights," (Washington, DC: White House, 2022).

⁴ Bureau of Arms Control, Verification and Compliance, "Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy," (Washington, DC: Department of State, 2023).

⁵ Anna Jobin, Marcello Lenca, and Effy Vayena, "The global landscape of AI ethics guidelines," *Nature Machine Intelligence* 1, (2019): 389–399, <https://doi-org.proxy.library.georgetown.edu/10.1038/s42256-019-0088-2>

⁶ Niels van Berkel, Eleftherios Papachristos, Anastasia Giachanou, Simo Hosio, and Mikael B. Skov. "A Systematic Assessment of National Artificial Intelligence Policies: Perspectives from the Nordics and Beyond," Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society, (October 26, 2020), <https://doi.org/10.1145/3419249.3420106>.

⁷ Anna Jobin, Marcello Lenca, and Effy Vayena, "The global landscape of AI ethics guidelines," *Nature Machine Intelligence* 1, (2019): 389–399, <https://doi-org.proxy.library.georgetown.edu/10.1038/s42256-019-0088-2>.

⁸ Zoe Stanley-Lockman, "Responsible and Ethical Military AI" (Center for Security and Emerging Technology, August 2021). <https://doi.org/10.51593/20200091>.

⁹ Niels van Berkel, Eleftherios Papachristos, Anastasia Giachanou, Simo Hosio, and Mikael B. Skov. "A Systematic Assessment of National Artificial Intelligence Policies: Perspectives from the Nordics and Beyond," Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society, (October 26, 2020), <https://doi.org/10.1145/3419249.3420106>.

¹⁰ National Institute of Standards and Technology, *AI Risk Management Framework: Second Draft* (Gaithersburg, MD: Department of Commerce, August 18, 2022), https://www.nist.gov/system/files/documents/2022/08/18/AI_RMF_2nd_draft.pdf.

¹¹ Organisation for Economic Co-operation and Development, *AI Principles-Overview*, Accessed June 13, 2023, <https://oecd.ai/en/ai-principles>.

¹² European Commission, *Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union legislative acts*, COM/2021/206 (Brussels: April 2021), <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206>

¹³ Sam Corbett-Davies, Emma Pierson, Avi Feller, and Sharad Goel. "A Computer Program Used for Bail and Sentencing Decisions Was Labeled Biased against Blacks. It's Actually Not That Clear." *The Washington Post*, October 17, 2016, <https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/>.

¹⁴ Organization for Economic Cooperation and Development, *Recommendation of the Council on Artificial Intelligence*, OECD/LEGAL/0449 (Paris: 2019), <https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449>.

¹⁵ Other examples of bilateral and multilateral agreements with similar goals and documents include the QUAD Principles on Technology Design, Development, Governance, and Use, AUKUS, France-Japan-Germany, Australia-Vietnam, the Canada-U.S. Innovation partnership, and the U.S.-UK Cooperation in AI R&D.

¹⁶ Australia Department of Industry, Science and Resources, *Australia's AI Ethics Principles* (Canberra: Australian Government, Accessed March 17, 2023), <https://www.industry.gov.au/publications/australias-artificial-intelligence-ethics-framework/australias-ai-ethics-principles>.

¹⁷ 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, version 2*, (2017). https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf.

¹⁸ Treasury Board of Canada Secretariat, *Directive on Automated Decision-Making*, (Ottawa: Government of Canada, 2019), <https://www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=32592>.

¹⁹ Canada, *Directive on Automated Decision-Making*.

²⁰ Government of Canada, *Responsible Use of Artificial Intelligence, Guiding Principles*, (Ottawa: Government of Canada, Accessed August 2023), <https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai.html#toc1>. Canada's Guiding Principles were also adopted by the Digital Nations, an international organization of digital countries who work to incorporate technology into government operations, at the D9 conference in 2018. At the time, the members of the D9 were Estonia, Israel, South Korea, New Zealand, the United Kingdom, Canada, Uruguay, Mexico, and Portugal.

²¹ Japan Council for Social Principles of Human-centric AI, *Social Principles of Human-Centric AI*, (2019), <https://ai.bsa.org/wp-content/uploads/2019/09/humancentricai.pdf>.

²² Japan Expert Group on How AI Principles Should be Implemented, *Integrated Innovation Strategy Promotion Council, Governance Guidelines for Implementation of AI Principles*, (2022), https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/pdf/20220128_2.pdf.

²³ The Governance Guidelines describe the objectives of action-targets as “general and objective ones that should be implemented by every AI company involved in AI business, typically the development/operation of AI systems that could have a certain level of negative impacts on society.”

²⁴ United Kingdom, *Data Protection Act*, (London: UK Government, 2018), <https://www.legislation.gov.uk/ukpga/2018/12/contents/enacted>.

²⁵ United Kingdom Information Commissioner's Office, *UK GDPR guidance and resources*, (London), <https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/>.

²⁶ United Kingdom Information Commissioner's Office and The Alan Turing Institute, *Explaining decisions made with AI*, (London: 17 October 2022), <https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/explaining-decisions-made-with-artificial-intelligence/>.

²⁷ United Kingdom Information Commissioner's Office, *Guidance on AI and data protection*, (London: Updated 15 March 2023), <https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/guidance-on-ai-and-data-protection/>.

²⁸ Another example of the influence of the UK's principles is the Secretary of State for Science, Innovation and Technology's Policy Paper *A pro-innovation approach to AI regulation*, Published March 29, 2023 and updated August 3, 2023.

²⁹ United States, Exec. Order No. 13960, 85 FR 78939 (2020).

³⁰ U.S., EO13960.

³¹ Exec. Order No. 13859, 84 FR 3967 (2019).

³² Office of Management and Budget (OMB), *OMB Memorandum M-21-06: Guidance for Regulation of Artificial Intelligence Applications*, (Washington, DC: White House, 2020).

³³ Office of Scientific and Technology Policy, *Blueprint for an AI Bill of Rights, About This Document*, (Washington, DC: White House, 2022), <https://www.whitehouse.gov/ostp/ai-bill-of-rights/about-this-document/>.

³⁴ Anna Jobin, Marcello Lenca, and Effy Vayena, “The global landscape of AI ethics guidelines,” *Nature Machine Intelligence* 1, (2019): 389–399, <https://doi-org.proxy.library.georgetown.edu/10.1038/s42256-019-0088-2>.

³⁵ Sample text analysis can be found in Appendix B.

³⁶ Australia, *Australia's AI Ethics Principles*.

³⁷ Canada, *Directive on Automated Decision-Making*.

³⁸ Japan, *Governance Guidelines*.

³⁹ Australia, *Australia's AI Ethics Principles*.

⁴⁰ Japan, *Social Principles of Human-Centric AI*.

⁴¹ UK ICO, *Guide to the UK GDPR: Accountability and Governance*.

⁴² For an interesting treatment of U.S. accountability options, see Zachary Arnold and Micah Musser, "The Next Frontier in AI Regulation Is Procedure," *Lawfare* (August 10, 2023).

⁴³ Australia, *Australia's AI Ethics Principles*.

⁴⁴ UK ICO, *Guide to the UK GDPR: Accountability and Governance*.

⁴⁵ UK ICO, *Guidance on AI and Data Protection*.

⁴⁶ UK ICO, *Guidance on AI and Data Protection*.

⁴⁷ Canada, *Directive on Automated Decision-Making*.

⁴⁸ UK ICO *Guidance on AI and data protection*.

⁴⁹ In the AI RMF 1.0, NIST defines interpretability as "the meaning of AI systems' output in the context of their designed functional purposes."

⁵⁰ Japan, *Social Principles of Human-Centric AI*.

⁵¹ Australia, *Australia's AI Ethics Principles*.

⁵² UK ICO and The Alan Turing Institute, *Explaining decisions made with AI*.

⁵³ U.S., EO13960.

⁵⁴ UK ICO and The Alan Turing Institute, *Explaining decisions made with AI*.

⁵⁵ UK ICO and The Alan Turing Institute, *Explaining decisions made with AI*.

⁵⁶ UK ICO and The Alan Turing Institute, *Explaining decisions made with AI*.

⁵⁷ Canada, *Responsible Use of Artificial Intelligence, Guiding Principles*.

⁵⁸ The type of disclosure required in Canada depends on the impact. Per the Directive on Automated Decision Making, “(Level 1 (decision will lead to impacts that are reversible and brief): can be provided in FAQ on a website, Level 2-4 (2: impacts likely reversible and short term, 3: difficult to reverse and ongoing, 4: impacts are irreversible and perpetual): in addition, meaningful explanation is provided with any decision that resulted in the denial of a benefit.”)

⁵⁹ U.S., EO13960.

⁶⁰ *Australia, Australia’s AI Ethics Principles*.

⁶¹ Japan, *Social Principles of Human-Centric AI*.

⁶² UK ICO and The Alan Turing Institute, *Explaining decisions made with AI*.

⁶³ NIST, AI RMF 1.0.

⁶⁴ *Australia, Australia’s AI Ethics Principles*.

⁶⁵ Japan, *Social Principles of Human-Centric AI*.

⁶⁶ UK ICO, *Guide to the GDPR, Principle (a)*.

⁶⁷ U.S. OMB, *Guidance for Regulation of Artificial Intelligence Applications*.

⁶⁸ *Australia, Australia’s AI Ethics Principles*.

⁶⁹ UK ICO, *Guide to the UK GDPR, Principle (a)*.

⁷⁰ *Australia, Australia’s AI Ethics Principles*.

⁷¹ UK ICO, *Guide to the UK GDPR, Principle (a)*.

⁷² The U.S. AI Bill of Rights goes a step further than the documents we textually analyzed by highlighting algorithmic discrimination protections. Algorithmic discrimination occurs when algorithms make “unjustified different treatment or impacts disfavoring people based on their race, color, ethnicity, sex (including pregnancy, childbirth, and related medical conditions, gender identity, intersex status, and

sexual orientation), religion, age, national origin, disability, veteran status, genetic information, or any other classification protected by law.”

⁷³ Japan, *Social Principles of Human-Centric AI*.

⁷⁴ Japan, *Social Principles of Human-Centric AI*.

⁷⁵ U.S. OMB, *Guidance for Regulation of Artificial Intelligence Applications*.

⁷⁶ Supreme Court of Canada, “*Baker v. Canada (Minister of Citizenship and Immigration)*,” [1999] 2 SCR 817.

⁷⁷ Immigration and Citizenship, *Procedural fairness*, (Ottawa: Government of Canada, Accessed 03-18-2023), <https://www.canada.ca/en/immigration-refugees-citizenship/corporate/publications-manuals/operational-bulletins-manuals/service-delivery/procedural-fairness.html#fair-impartial-decision-making>.

⁷⁸ U.S. OMB, *Guidance for Regulation of Artificial Intelligence Applications*.

⁷⁹ Federal Bureau of Investigation, *The China Threat*, (Washington, DC: FBI, Accessed August 10, 2023), <https://www.fbi.gov/investigate/counterintelligence/the-china-threat>.

⁸⁰ The White House, “The United States, Joined by Allies and Partners, Attributes Malicious Cyber Activity and Irresponsible State Behavior to the People’s Republic of China,” (Washington, DC: The White House, July 19, 2021), <https://www.whitehouse.gov/briefing-room/statements-releases/2021/07/19/the-united-states-joined-by-allies-and-partners-attributes-malicious-cyber-activity-and-irresponsible-state-behavior-to-the-peoples-republic-of-china/>.

⁸¹ UK GDPR, Article 5(1)(c).

⁸² Japan, *Social Principles of Human-Centric AI*.

⁸³ UK ICO, *Guidance on AI and data protection*.

⁸⁴ U.S. OMB, *Guidance for Regulation of Artificial Intelligence Applications*

⁸⁵ Japan, *Social Principles of Human-Centric AI*.

⁸⁶ Australia, *Australia’s AI Ethics Principles*.

⁸⁷ Australia, *Australia’s AI Ethics Principles*.

⁸⁸ U.S. OMB, *Guidance for Regulation of Artificial Intelligence Applications*.

- ⁸⁹ UK ICO, *Guidance on AI and Data Protection*.
- ⁹⁰ NIST, *AI RMF 1.0*.
- ⁹¹ Australia Government, *Australia's AI Ethics Principles*.
- ⁹² Canada, *Directive on Automated Decision-Making*.
- ⁹³ Japan, *Governance Guidelines for Implementation of AI Principles*.
- ⁹⁴ UK ICO, *Guide to the UK GDPR, Principle (a)*.
- ⁹⁵ U.S. OMB, *Guidance for Regulation of Artificial Intelligence Applications*.
- ⁹⁶ U.S. OMB, *Guidance for Regulation of Artificial Intelligence Applications*.
- ⁹⁷ Canada, *Directive on Automated Decision-Making*.
- ⁹⁸ U.S., EO13960.
- ⁹⁹ Canada, *Responsible Use of Artificial Intelligence, Guiding Principles*.
- ¹⁰⁰ Canada, *Responsible Use of Artificial Intelligence, Guiding Principles*.
- ¹⁰¹ Deborah Raji et al., "The Fallacy of AI Functionality," *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*, (June 2022): 959–972, <https://doi.org/10.1145/3531146.3533158>.
- ¹⁰² See NIST's list of key characteristics in the AI RMF 1.0 and ISO/IEC TS 5723:2022
- ¹⁰³ Commonwealth Ombudsman Office, "Lessons learnt about digital transformation and public administration: Centrelink's online compliance intervention," July 2017, https://www.ombudsman.gov.au/__data/assets/pdf_file/0024/48813/AIAL-OCI-Speech-and-Paper.pdf.
- ¹⁰⁴ Martineau, P. "Toronto Tapped Artificial Intelligence to Warn Swimmers. The Experiment Failed," *The Information*, November 4, 2022, <https://www.theinformation.com/articles/when-artificial-intelligence-isnt-smarter>.
- ¹⁰⁵ E. Constantaras, G. Geiger, J. Braun, D. Mehrotra, and H. Aung, "Inside the Suspicion Machine," *Wired* March 6, 2023, <https://www.wired.com/story/welfare-state-algorithms/>.
- ¹⁰⁶ Canada, *Directive on Automated Decision-Making*.

- ¹⁰⁷ Australia, *Australia's AI Ethics Principles*.
- ¹⁰⁸ Japan, *Governance Guidelines for Implementation of AI Principles*.
- ¹⁰⁹ U.S. OMB, *Guidance for Regulation of Artificial Intelligence Applications*.
- ¹¹⁰ ISO/IEC TS 5723:2022(en) "Trustworthiness — Vocabulary," <https://www.iso.org/obp/ui/#iso:std:iso-iec:ts:5723:ed-1:v1:en>.
- ¹¹¹ UK ICO and The Alan Turing Institute, "What goes into an explanation?" (London: Information Commissioner's Office, 17 October 2022), <https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/explaining-decisions-made-with-artificial-intelligence/part-1-the-basics-of-explaining-ai/what-goes-into-an-explanation/>.
- ¹¹² U.S. OMB, *Guidance for Regulation of Artificial Intelligence Applications*.
- ¹¹³ Australia, *Australia's AI Ethics Principles*.
- ¹¹⁴ Canada, *Directive on Automated Decision-Making*.
- ¹¹⁵ ISO/IEC TS 5723:2022.
- ¹¹⁶ Andrew J. Lohn, "Estimating the Brittleness of AI: Safety Integrity Levels and the Need for Testing Out-Of-Distribution Performance." Arxiv, September 2, 2020, <https://doi.org/doi.org/2009.00802>.
- ¹¹⁷ Canada, *Directive on Automated Decision-Making*.
- ¹¹⁸ Japan, *Social Principles of Human-Centric AI*.
- ¹¹⁹ UK ICO, *Guidance on AI and Data Protection*.
- ¹²⁰ U.S., EO13960.
- ¹²¹ Australia, *Australia's AI Ethics Principles*.
- ¹²² D. Danks and D. Trusilo The Challenge of Ethical Interoperability. *DISO* 1, 11 (2022), <https://doi.org/10.1007/s44206-022-00014-2>.
- ¹²³ OECD.AI (2021), powered by EC/OECD (2021), database of national AI policies, <https://oecd.ai/en/>.
- ¹²⁴ Emerging Technology Observatory, "Country Activity Tracker: Artificial Intelligence," <https://cat.eto.tech/>.