

Trust Issues: Discrepancies in Trustworthy AI Keywords Use in Policy and Research

Autumn Toney-Wails
Georgetown Univeristy
Washington D.C., United States
autumn.toney@georgetown.edu

Kathleen Curlee
Georgetown Univeristy
Washington D.C., United States
kathleen.curlee@georgetown.edu

Emelia Probasco
Georgetown Univeristy
Washington D.C., United States
emelia.probasco@georgetown.edu

ABSTRACT

How governments, practitioners, and researchers define artificial intelligence (AI) ethics significantly impacts the AI models and systems designed and deployed. Thus, the convergence of policy goals and technical approaches is necessary for international norms and standards on trustworthy AI. Defining, much less achieving trustworthy AI characteristics, however, entails clear communication through consensus on the meaning of field-specific terms. This paper presents an analysis of over 322,000 scientific research papers and the national documents from five countries (Australia, Canada, Japan, the United Kingdom, and the United States) on trustworthy AI in order to provide an in-depth review and comprehensive understanding of the similarities and differences between governments' and researchers' definitions and frameworks. While we identified substantive and relevant differences among policy documents and scientific research, the differences do not represent substantial disagreements among the common principles for trustworthy AI terms. Overall we found broad agreement across documents' trustworthy AI term use, suggesting that nuanced differences could be overcome in an effort to create more global policies and aligned research.

CCS CONCEPTS

• **Social and professional topics** → **Computing / technology policy**.

KEYWORDS

Trustworthy AI, National AI Policy, Scientific Research

ACM Reference Format:

Autumn Toney-Wails, Kathleen Curlee, and Emelia Probasco. 2024. Trust Issues: Discrepancies in Trustworthy AI Keywords Use in Policy and Research. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*, June 03–06, 2024, Rio de Janeiro, Brazil. ACM, New York, NY, USA, 12 pages. <https://doi.org/10.1145/3630106.3659035>

1 INTRODUCTION

The development of trustworthy artificial intelligence (AI) and machine learning (ML) systems impacting society requires effective international regulations containing measurable, standardized

criteria. In addition to judgements made by policymakers, these regulations should include input from the scientific community who can better align legislation with state-of-the-art research to develop definitions, techniques, and tools used to assess trustworthiness. However, collaborative efforts demand considerable time from the policy and scientific communities and are logistically challenging to coordinate on an international scale. Additionally, global scientific research is published at a significantly higher rate than national policies are formed, exaggerating discrepancies in the key terms and definitions used in regulations and research.

Policy and research communities are actively engaged in focused efforts to define and measure trustworthy AI, but do so inconsistently among themselves and across nations, as they are racing to keep the pace of AI/ML system applications. It is challenging to evaluate the effectiveness, consistencies, and discrepancies in trustworthy AI from a policy or research perspective, as there is no international, standardized framework or taxonomy for assessing the trustworthiness of an AI system. Prior work has analyzed national policies (individually [4, 29, 32, 33] and collectively [9, 12, 30, 34]) and presented reviews and analysis on research focused on trustworthy AI topics [2, 7, 15, 31]. However, due to the broadness of the topic, most work is either narrowly scoped or lacks comprehensive detail from manual review of numerous policy and research documents.

In an effort to evaluate the similarities and differences between national policies and scientific research, we present an empirical and quantitative comparative analysis on trustworthy AI policies from five countries (spanning four continents) and 322,209 scientific publications focused on trustworthy AI-related research. We select national policies from Australia, Canada, Japan, the United Kingdom, and the United States, as all five countries adhere to the Organisation for Economic Co-operation and Development (OECD) Recommendation of the Council on Artificial Intelligence [17] and published specific national guidance on the governance of AI. For scientific publications, we use a merged corpus of scholarly literature (comprising of six databases) and leverage the National Institute of Standards and Technologies' (NIST) AI Risk Management Framework v1.0 for a set of trustworthy AI terms to identify relevant publications. We manually review and summarize our findings on 10 national policy documents and 650 research publications.

Our findings highlight that while both national regulations and scientific research overlap some of their trustworthy AI key terms use, definitions differ but not excessively. The terms most used in policy documents are different from those used in scientific research; for example, *accountability* is one of the most frequently appearing terms in policy documents, whereas it is the least used term in research publications. While trustworthy AI policy and

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

FAccT '24, June 03–06, 2024, Rio de Janeiro, Brazil

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 979-8-4007-0450-5/24/06

<https://doi.org/10.1145/3630106.3659035>

research is evolving, our analysis highlights that discrepancies in definitions can be aligned with input from both communities, as the gaps are not overwhelming. However, there is a more noticeable gap in the frequency of term mentions between policy documents and research, suggesting that the policy and research communities are focused on different aspects on trustworthy AI. In the following sections we describe the datasets used for analysis (Section 2), define the criteria for identifying trustworthy AI documents (Section 3), present our empirical and quantitative results (Section 4), and provide a discussion on our findings (Section 5). Sections 2-4 are organized in two distinct subsections, with the first discussing the policy documents and the second discussing the scientific research.

2 AI POLICY AND RESEARCH DATASETS

In this work, we analyze trustworthy AI-related national policy documents and scientific research publications, both restricted to English-language. Section 2.1 lists the policy documents selected for analysis and Appendix B contains brief descriptions of the documents. Section 2.2 describes the curation of our scientific research corpus.

2.1 National Trustworthy AI Documents

While each country in our analysis has designed and published numerous AI strategy, governance, or policy documents written by different agencies across each government, we prioritize documents that appear most likely to influence future AI policies. We select policy documents by the following three criteria: 1) published by an institution charged with the highest level of guidance and governance within that nation, 2) provided guidance regarding the development and/or use of AI by the government or by private entities, and 3) included specific references to ethical or trustworthy AI terms and principles in a substantial way. While a country may have multiple trustworthy AI-related policy documents, providing a full inventory is outside the scope of this paper. The policy documents selected for analysis vary from voluntary guidelines to binding national policy, as not all countries have implemented legally binding AI policy. We select the following documents from each country for analysis:

Australia: AI Ethics Principles [5], **Canada:** Directive on Automated Decision-Making [19]; Responsible Use of Artificial Intelligence Guiding Principles [21], **Japan:** Social Principles of Human-Centric AI [10]; Governance Guidelines for Implementation of AI Principles [27], **United Kingdom:** ICO Guide to the UK General Data Protection Regulation and Data Protection Act[14]; ICO and Alan Turing Institute Explaining Decisions Made with AI [26]; ICO Guide on AI and Data Protection [24], **United States:** Presidential Executive Order 13960 [23]; White House OMB Guidance for Regulation of Artificial Intelligence Applications [36].

2.2 Scientific Research Documents

We use a combined corpus of scientific research containing documents from Digital Science Dimensions¹, Clarivate’s Web of Science,

Microsoft Academic Graph, China National Knowledge Infrastructure², arXiv, and Papers With Code. We follow a two-step process to deduplicate documents across all databases, as there is not a singular common publication identifier (e.g., DOI). First, we select six document identifiers that are not all independently unique: 1) DOI, 2) citations, 3) normalized abstract, 4) normalized author names, 5) normalized title, and 6) publication year. We normalize text features using the Normalization Form Compatibility Composition standard, which decomposes unicode characters by compatibility, recomposes them by canonical equivalence, de-accent letters, strips copyright signs, HTML tags, punctuation, non-alphanumeric characters, and numbers, and removes white space from the strings. If any three of the six document features match to another document, we assign those documents to a merged ID. Second, we apply the SimHash fuzzy matching algorithm (using a rolling window of three characters) on the articles that were not deduplicated in our first step, comparing documents’ titles and abstracts that have the same publication year [16]. Any documents matched in this second step also receive a merged ID and are considered deduplicated. The remaining documents with no matches found in step one or two are considered to be unique documents and merged into the final corpus used in analysis, resulting in 184,381,319 documents.

3 IDENTIFYING TRUSTWORTHY AI POLICY AND RESEARCH

For each document type (policy and scientific publication), we define our criteria for trustworthy AI identification. Policy documents were identified in Section 2.1, but required manual annotation for key term mentions. In contrast, we use a set of pre-defined key terms to identify relevant scientific publications for this analysis.

3.1 AI Policy Document Analytical Framework

For each policy document, the authors manually annotated common terms and explanations related to AI trustworthiness, using the text analysis platform Dedoose³, to surface a set of trustworthy AI terms listed in Table 1. Our annotations identified terms used directly in the policy documents examined, as opposed to searching for key terms constructed independently from prior work or other policy documents. Thus, the terms in Table 1 do not necessarily reflect the most salient trustworthy AI terms for global trustworthy AI policy, but rather the most salient trustworthy AI terms in the national policies of Australia, Canada, Japan, the UK, and the U.S.

Figure 1 provides two examples of policy document annotation, one from Japan’s Social Principles of Human-Centric AI and one from Australia’s AI Ethics Principles [5, 10]. These examples illustrate the relative frequency and use of the trustworthy AI term set used for annotation. These annotations surfaced the most commonly used terms, as well as identified specific use-cases of commonality and variance across the five countries in our analysis.

¹Data sourced from Dimensions, an inter-linked research information system provided by Digital Science <http://www.dimensions.ai>

²All China National Knowledge Infrastructure content is furnished for use in the United States by East View Information Services, Minneapolis, MN, USA

³<https://www.dedoose.com/>

Table 1: Trustworthy AI themes used for policy document annotation with their corresponding description.

Term	Description
Accessibility	Any references to the AI system being accessible, adhering to accessibility guidelines, or enhancing accessibility
Accountability	Any references to the AI being made accountable
AI Lifecycle	Any references to any part of the AI lifecycle
AI System Design	Any references to the design or creation of the AI system or how the system works
Discrimination	Any references to discrimination, whether ameliorating or avoiding it as well as references to protected classes
Explainability	Any references to the AI being explainable, either in reference to the outcomes of the AI system or the system itself
Fairness	Any references relating to fairness and its traditional conceptions, including equality and unfairness
Human Rights	Any references to human rights, whether enabling or not inhibiting
Inclusivity	References to the AI being inclusive
Law	References to the law in any form
Outcomes	References to the outcomes of an AI system
People/Individuals	References to groups of people, individuals, or actions by people
Privacy	Any references to an AI system protecting privacy
Procedural Fairness	References to or aspects of procedural fairness (distinguished from general fairness)
Security	References to an AI system being secure
Statistics	References to statistics or statistical analysis in relation to an AI system
Transparency	Any references to an AI system being transparent
Unjust/Unlawful	Any references to an AI system being unjust or unlawful
Users	Any references to creators of an AI system, researchers, or end-users (distinguished from people/individuals)

1. Accessibility 2. AI Lifecycle 3. AI System Design 4. Discrimination
5. Fairness 6. Human Rights 7. Inclusivity 8. People/Individuals 9. Unjust/Unlawful 10. Users

Japan:

“Under AI’s design³ concept, all people⁸ are treated fairly⁵ without unjustified⁹ discrimination⁴ on the grounds of diverse⁶ backgrounds such as race, gender, nationality, age, political beliefs, religion⁶, and so on.”

Australia:

“Throughout their lifecycle², AI systems should be inclusive⁷ and accessible, and should not involve or result in unfair⁵ discrimination⁴ against individuals, communities or groups⁸. This principle aims to ensure that AI systems are fair⁵ and that they enable inclusion⁷ throughout their entire lifecycle². AI systems should be user-centric¹⁰ and designed³ in a way that allows all people interacting with it⁸ to access the related products or services¹. This includes both appropriate consultation with stakeholders⁸ who may be affected by the AI system throughout its lifecycle², and ensuring people⁸ receive equitable access and treatment⁵.”

Figure 1: Example annotations from Japan’s Social Principles of Human-Centric AI and Australia’s AI Ethics Principles of sections that discuss fairness. Terms and phrases are both color-coded and numbered according to the term they map to.

3.2 Trustworthy AI Scientific Publications

To identify trustworthy AI research publications, we first identify AI-related publications and then select a subset that contain trustworthy AI terms. We filter for documents in a 10-year window, being published between 2010 and 2021, to scope our publication set on recent and relevant papers. We apply an AI classifier to identify AI-related publications; classifier labels are binary (AI and non-AI) [8]. The AI classifier uses arXiv publications as training

data, as arXiv publications receive author assigned topic labels, and the SciBERT language model trained on Semantic Scholar [6]. The following arXiv labels were considered as positive class (AI) and all other arXiv publications were considered negative class (non-AI): cs.AI, cs.CL, cs.CV, cs.LG, stat.ML, cs.MA, cs.RO. Using this AI classifier we identify 2,324,124 AI-related publications in our scientific document corpus.

Drawing on the National Institute of Standards and Technologies’ AI Risk Management Framework v1.0 (NIST AI RMF), we use a

set of trustworthy AI terms to identify trustworthy AI-related publications from the AI publication set:

- Accountability/Accountable
- Bias
- Explainability/Explainable
- Fairness
- Interpretability/Interpretable
- Reliability, Reliable
- Resilience
- Robustness
- Safe, Safety
- Secure, Security
- Transparency
- Trust

Our list deviates slightly from the term list provided in the NIST AI RMF v1.0. Officially, NIST uses the terms bias-managed and privacy-enhanced, but we selected the simplified “bias”. Robustness is not explicitly listed, but rather defined within NIST’s discussion of valid and reliable. We include the term “trust” as it is in the overarching theme. We exclude valid and accurate as these terms are far more general and vague to use as a trustworthy AI term identifiers. Searching through publications’ titles and abstracts we identify a set of 322,209 trustworthy AI-related publications that had at least one trustworthy AI term mention and were classified as AI-related.

4 TRUSTWORTHY AI DOCUMENT ANALYSIS

The following sections describe the empirical and quantitative findings from our manual review and semantic search analysis. Section 4.1 provides summaries of our manual review of national policy documents and Section 4.2 provides term frequency and use analysis in the scientific publications.

4.1 National Policy Documents

After manual review, six terms appeared consistently in the high-level governmental guidance documents we examined: accountability, explainability, fairness, privacy, security, and transparency. This finding is consistent with prior work that identified responsibility, transparency, justice and fairness, privacy, and non-maleficence to be common terms [12]. With these six key terms, our policy analysis focuses on how the terms are used in each country’s key documents. For each section below, we define the key term and include surfaced subtopics and their descriptions. Figure 2 displays the topics defined, and denotes the countries that included the corresponding topic in their national AI policy documents.

4.1.1 Accountability. One consistent theme among national policy discussion of accountability was that humans must be accountable for the adverse outcomes of AI systems for which they bear some responsibility. Nations vary 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. We identify three main differences 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. Australia’s policy mentions an expectation that “human oversight of AI systems should be enabled” and that organizations must “consider the appropriate level of human control or oversight for the particular AI system or use case” [5]. 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. 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”[10]. 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 side-step 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: Australia, Japan, and the UK all note the person affected by the AI must be part of any accountability processes [5, 10, 24]. The centrality of the affected person in an accountability process is echoed in each country’s conception of fairness. Australia and the UK specifically emphasize the need for affected persons to be able to challenge an AI’s decisions. The UK ICO is more specific, stating that processes and results must be documented to an “auditable standard” for accountability [25]. Additionally, Australia includes the potential for compensation and a timely accountability process for those harmed by an AI [5].

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 [19]. 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 [24]. Data protection officers are also accountable for understanding the GDPR and its impact on AI tools and systems.

4.1.2 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. 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. Specifications on the audiences for AI explanations can guide developers to create systems designed for those audiences and their circumstances.

Country expectations for audiences vary significantly; Japan is the most limited in its audience expectations, stating simply that explanations should be provided on a “case by case basis” [10]. In contrast, Australia’s principles include specified affected audiences: users, creators, legal representatives, and the public [5]. 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 [25]. The U.S. issues a blanket statement for






Topic	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			✓		✓

Figure 2: Trustworthy AI topic discussion across all national policy documents. Check marks denote the topics that were discussed in the corresponding country's national policy documents.

explanations to “others, as appropriate” [23]. 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.

Affected Users: All countries except 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 [14]. 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” [24].

What to Explain: Notification, System Structure, and System Outcomes: The five countries generally recognize three points that require explanation: 1) 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), 2) an explanation of how the system works, and 3) an explanation of the system’s output or decision. The UK and Canada embrace all three points for explanation: notification, system structure, and system outcomes. Specifically, 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 [26]. Canada adds to this list a requirement to explain the training data for the system and, if applicable, the way it was collected [21]. Both Canada and the UK further delineate expectations for explainability based on the impact level of an AI system. In Canada, AI systems that have reversible and brief impacts have a lower expectation for explainability than systems that

have irreversible or perpetual impacts. 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 and Australia and Japan mention a need for the explainability of system outcomes or results, but not necessarily how the system works.

Specific Guidance on Explainable Approaches: Unlike the other nations, the UK's guidance on explaining AI details 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) [26].

4.1.3 Fairness. Fairness was consistently emphasized 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 [1]. Additionally, fairness is a difficult principle to define mathematically, morally, or politically. Here, we note the similarities among the five countries, especially the importance of engaging an affected user in a process of determining fairness and preventing discrimination.

Role of Affected User: Australia, Japan, the UK, and the U.S. 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.

Importance of Disclosure or Consent: Australia and the UK provide particular clarity on the need to elicit informed consent from users who may interact with an AI system (echoed in their discussions of 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, Japan, the UK, and the U.S.; however, only Japan mentions specific categories that should be protected (e.g., age, gender, nationality, race, and religion). While the other countries do not list specific protected classes, they instead emphasize inclusiveness and accessibility. Of note, Japan and the U.S. emphasize the democratic notions of civil rights or civil liberties in their definitions of fairness [10, 36].

Procedural Fairness: Canada is unique in that it does not discuss discrimination, but instead draws upon its established concept of procedural fairness. 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" [19]. 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. This notion of procedural fairness informs Canada's approach to transparency and explainability as well.

4.1.4 Privacy. All five countries include the term privacy in their policy documents, with many references to established guidelines on data protection. There are notable differences in their statements which we summarize as what is to be protected (should it include intellectual property?), how is it to be protected (by security or by a data minimization standard?), and why is it protected (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 [36]. This may be linked to the U.S. assumption that safeguarding intellectual property is foundational to economic growth or to America's vocal concerns about the theft of intellectual property by China [11]. While American allies also recognize the importance of intellectual property no other national policy documents mentioned IP explicitly in their descriptions of privacy and AI.

Data Minimization: The UK highlights data minimization as a method for enhancing privacy. It notes that "personal data shall be adequate, relevant, and limited to what is necessary in relation to the purposes for which they are processed" [24]. 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 [14].

Privacy and Democratic Values: Japan, the UK, and the U.S. link privacy to individual rights and freedoms, with the U.S. documents having the most mentions. In EO13960, the phrase "privacy, civil rights, civil liberties" is used five times, and in two instances the phrase "American values" is included [23]. 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" [10]. Other countries mention democratic values, but not as an aspect of privacy (e.g., Australia highlights democratic values as a component of its "human-centered values" principle) [5].

4.1.5 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 (Australia, the UK, and the U.S.) of the five countries include guidance to build and operate systems in a way that fortifies them against an attack.

Risk Management Approach: All five countries explicitly address security concerns through risk assessment or risk management frameworks and processes. The documents acknowledge that AI systems contain risk and that governance is a process of managing risk.

Preparing for Attack: Australia, the UK, and the U.S. all show concern for malicious attacks against AI systems. Australia and the U.S. specifically mention the requirement for resilience (i.e., 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" [14].

4.1.6 Transparency. Transparency is related to, but distinct from, explainability. The NIST AI RMF v1.0 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” [1]. 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: All five countries 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” [36]. While related to the previously discussed principles of explainability and fairness, in this instance all countries agree on necessary disclosure as a part of the principle of transparency. Canada’s approach to providing notice is different, as 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.

Balancing Transparency with Privacy: The U.S. and Canada 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” [21].

4.2 Scientific Research Key Term Mentions

First, we analyze publication output over time, to identify if AI-related papers that mention trustworthy AI terms follows the same publication output trend as all AI-related publications. We compute the yearly percent change in publication output for the AI-related publications (~2.3M in total) and trustworthy AI-related publications (~320K in total), shown in Figure 3. The yearly percentage change trends are similar, with the trustworthy AI publications outpacing AI publication output, relative to their respective set sizes. This result suggests that the NIST trustworthy AI terms are relevant to AI researchers, and are increasingly being used at a faster rate than general AI publication output.

We count the occurrence of each trustworthy AI term in our trustworthy AI publication set—searching over titles and abstracts—and present the results in Figure 4. Unique terms are counted per document, as opposed to per mention (e.g., if a title and abstract mention fairness more than once, fairness is only counted once). If a publication mentions multiple terms, each term receives a count of one (e.g., if a publication mentions safety and privacy, both safety and privacy receive a count of one). Reliability, robustness, safety, security, and bias are the top five most frequently appearing

terms, with accountability, resiliency, transparency, fairness, and explainability being the least frequently appearing trustworthy AI terms. There is a wide range of term counts, specifically in 2021 reliability has 14,812 mentions (highest frequency) whereas accountability has 276 mentions (lowest frequency). This initial comparison provides an understanding of the representation of the trustworthy AI terms (curated by NIST) in scientific research publications.

Next, we compute the pair and triple sets of co-occurring terms to identify commonly co-occurring trustworthy AI terms, displayed in Table 2. Of the 332,209 trustworthy AI keyword publications, 12% mention more than one of the trustworthy AI terms. For two-term co-occurrence pairs, we find (reliability, robustness) and (reliability, safety) with the highest frequencies, 4,500 and 4,195 respectively. For the three-term co-occurrence triples, we find (reliability, safety, security) and (reliability, robustness, and safety) with the highest frequencies, 286 and 274 respectively. While security has the fourth highest frequency in Figure 4, Table 2 highlights its wide use in conjunction with other trustworthy AI terms, as it appears in three of the two-term pairs and in four of the three-term triples.

Table 2: Trustworthy AI 2-term and 3-term co-occurrence sets in publication titles and abstracts.

2-Term		3-Term	
Terms	Freq	Terms	Freq
reliability, robustness	4,500	reliability, safety, security	286
reliability, safety	4,195	reliability, robustness, safety	274
safety, security	2,775	privacy, security, trust	229
privacy, security	2,621	privacy, safety, security	177
reliability, security	2,223	reliability, robustness, security	150

While, the key term frequency analysis is useful, it does not provide the context in how the terms are being used in research. In order to provide a contextual understanding of how the trustworthy AI terms are being referenced in research publications, we manually review the 50 most cited papers for each term in 2021 (650 publications total).

4.2.1 Reliability. As the most frequently mentioned trustworthy AI term, reliability may appear the most due to its wide range of use. Over half of the publications that mentioned reliability used the term consistently with NIST’s definition: “the ability of an item to perform as required without failure, for a given time interval, under given conditions” [1]. The publications that used reliability in alignment with NIST’s definition were asserting or documenting the reliability of a specific approach or application of AI/ML. For example, abstracts frequently included the phrase “our method produces reliable results,” or “we confirmed the reliability of our models.” Approximately one-third of the publications, were concerned with the implementation of AI to improve the reliability of non-AI systems (e.g., the reliability of COVID-19 detection or the reliability of a tunnel-boring machine), and thus used the term reliability but not with respect to AI. This use case highlights that researchers use the term with respect to the promise of AI to improve

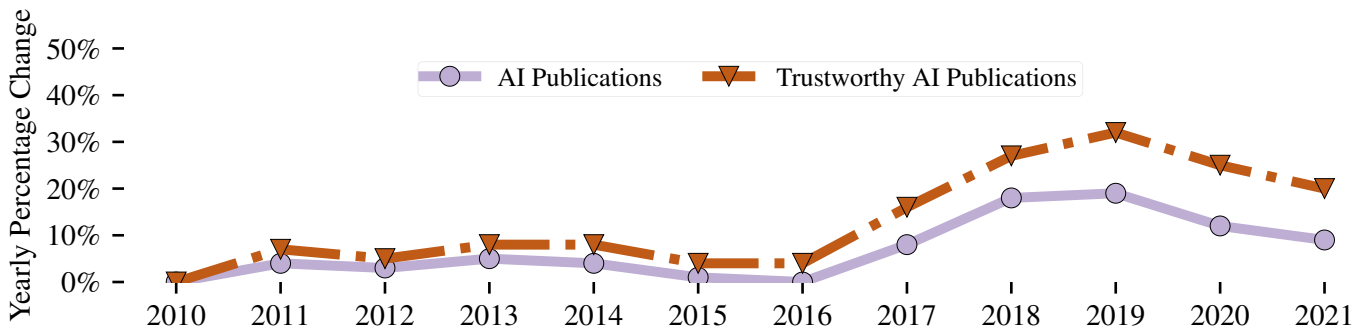


Figure 3: AI and trustworthy AI yearly percentage change in publication output between 2010 and 2021.

Number of Publications

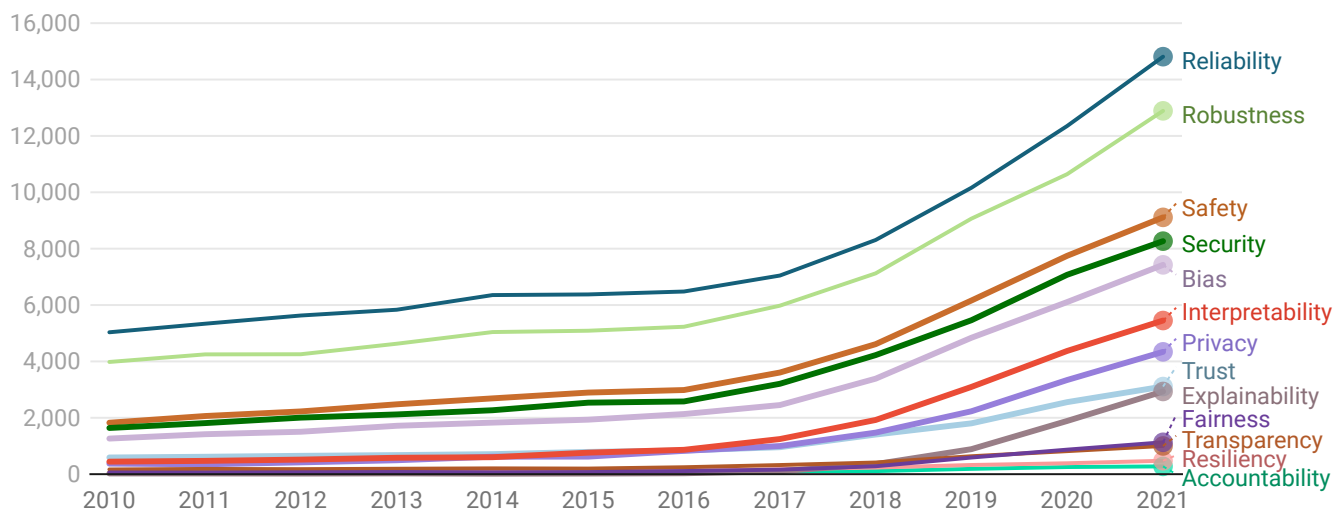


Figure 4: Trustworthy AI term frequency over time, with frequencies determined by mentions in publications’ titles and abstracts.

the reliability of current, non-AI systems, which does not align with the policy-framed definition of making AI systems reliable.

4.2.2 *Robustness.* The NIST and International Standards Organization (ISO) define robustness as the “ability of a system to maintain its level of performance under a variety of circumstances”[1, 28]. Of the 50 in this set, the majority mention robustness in alignment with this definition, specifically using robustness as an evaluation metric for a particular proposal (i.e., measuring the robustness of a federated learning approach). In contrast to other term usages, assertions of robustness were frequently in relation to that of another method or approach, indicating that measures of robustness were perceived as relative rather than absolute.

4.2.3 *Safety.* Referencing the ISO guidance, NIST specifies that the characteristic of safety requires that AI systems not, “under defined conditions, lead to a state in which human life, health, property, or the environment is endangered”[1, 28]. Approximately half of the 50 papers reviewed in this set used safety in alignment with NIST. The remaining publications focused on implementing

AI to improve the safety of a current process or technology (e.g., in construction, medicine, or manufacturing). One-third of the 50 papers were concerned with the application of AI in autonomous vehicles (mostly cars but also seagoing vessels). There was notable uses of safety in connection with the security or privacy of personal data, with several papers mentioning other key terms, including robustness, reliability, and explainability.

4.2.4 *Security.* As the fifth most frequently used term in research, security did not align with NIST’s definition as much as other terms. NIST defines security as “AI systems that can maintain confidentiality, integrity, and availability through protection mechanisms that prevent unauthorized access and use” [1]. Security was most frequently used as a reference to the application of AI in security research (cybersecurity, security in IoT). Eight of the 50 papers used security in a tangential reference; for example, food security as a motivation for improving a crop monitoring algorithm. The remaining few publications mentioned security in a similar use case as defined by NIST (e.g., assessments of the relative security

of a particular AI method and proposals to improve the security of an AI-enabled system).

4.2.5 Bias. NIST accounts for a broad definition of bias in its AI RMF, pointing out that “bias is not always a negative phenomenon,” namely when discussed in a technical sense [1]. Albeit, NIST connects bias to fairness and is most concerned that bias in AI not create, perpetuate, or amplify harm to individuals. But the appearance of the term bias in AI research is not necessarily focused on harmful discrimination. Rather, the term reflects NIST’s broader statement that bias is not necessarily harmful and that it is in fact an essential part of building an AI algorithm. For example, nearly two-thirds of the 50 abstracts examined referenced the role or need for “inductive bias”, or the “weights and biases” necessary to develop an algorithm. The remaining third of the titles and abstracts using the word bias addressed racial or gender bias or methods for addressing bias through techniques connected to other keywords such as explainability.

4.2.6 Interpretability and Explainability. We discuss interpretability and explainability together, as they are intimately connected and can be challenging for non-specialists to separate in the research literature. NIST makes a clear distinction between the need to properly interpret the recommendation of an AI system (interpretability) and the related need to represent “the mechanisms underlying AI systems’ operation” (explainability) [1]. While NIST clearly distinguishes between the definitions of the two terms, these definitions are notoriously not universally accepted. For example, Amazon’s definitions of the two terms are nearly reversed from NIST’s [13], and several of the top 50 cited publications use explainability and interpretability synonymously, highlighting misalignment among researchers as well [35].

Notably, the research area of Explainable AI (XAI) focuses on designing AI systems that an end-user can trust, with both interpretability and explainability as critical components. The majority of the 50 publications reviewed here used the term explainable in the XAI phrase, and explainability to reference the ability to identify how the model makes decisions. Most publications that mention the term interpretability did so to describe the evaluation of or improvement to deep learning classification outputs, specifically in regards to an XAI framework. Several publications surveyed deep learning models that either lacked interpretability or asserted the state-of-the-art for interpretable outputs.

4.2.7 Privacy. The vast majority of the 50 top-cited trustworthy AI papers that included the word privacy were aligned with the NIST AI RMF and referenced techniques and approaches to improve the privacy of user or device data, often for specific applications as opposed to general frameworks. Among the publications concerned with improving privacy, the majority discussed either the potential for federated learning to improve privacy or ways to improve federated learning approaches to minimize the loss of performance observed as a tradeoff for the technique. Overall, one-third of the papers focused on privacy issues for IoT, and slightly less than one third focused on privacy in the medical field. This may be attributed in part to a high concern over the sharing of data in the midst of the COVID-19 pandemic (10 papers specifically mentioned COVID-19).

4.2.8 Trust. NIST’s AI RMF is concerned on the whole with the creation of trustworthy AI, and thus, defines trustworthy AI by its key characteristics. Given this overarching concern with the trustworthiness of AI systems, we specifically include the term trust in our analysis. We find that the majority of trust mentions co-occur with another trustworthy AI term, particularly privacy, security, and explainability (40 of the 50 papers). Of the publications with no co-occurring trustworthy AI term, the research was focused on the trustworthiness of an AI system as NIST would consider it.

4.2.9 Fairness. The 50 most cited publications mentioning fairness are mainly aligned with NIST’s definition, and echo NIST’s assertion that fairness is a socio-technical issue that can vary across groups or cultures [1]. Publications in this set included those that strongly linked fairness to bias and discrimination, explored fairness beyond bias or discrimination, examined definitions of fairness, and focused on technical fairness (similar to the technical bias publications). In addition, several publications explicitly examined or evaluated the fairness of a particular application of AI/ML.

4.2.10 Transparency. NIST’s AI RMF defined transparency as ensuring “information about an AI system and its outputs is available to individuals interacting with such a system” [1]. The majority of publications that mentioned transparency aligned with the NIST definition, with titles and abstracts that also included another one of NIST’s key terms, in particular explainability and/or interpretability. Several of the publications in this set discussed the relative merits of post-hoc explainability for transparency, as well as “transparent algorithms.” Another subgroup of the publications detailed the importance of transparency to trust and/or the adoption of AI/ML. Several of the papers included specific proposals to improve AI transparency for a given use or to improve the transparency of datasets.

4.2.11 Resiliency. Two-thirds of the 50 titles and abstracts mentioning resiliency did so in alignment with NIST (“withstand[ing] unexpected changes in their environment or use”) [1]. Publications mentioned resiliency as an evaluation metric, made proposals for general resiliency approaches such as digital twins, and voiced concerns about fault tolerance as a component of resilience. Similar to safety and security, though to a lesser extent, approximately one-third of the top 50 papers were not about resilience as a characteristic of AI, but rather about the application of AI to improve resilience in a non-AI context. Research topics in this publication set covered a wide range of resiliency use cases; for example, resiliency in the face of climate change, the resilience of robots to environmental shifts, supply chain resilience, and even the application of AI to monitor pigs for indications of animal resiliency.

4.2.12 Accountability. Unlike the other trustworthy AI terms, the majority of publications that mentioned accountability were studies of the concept or importance of accountability, and not specific proposals to improve or address accountability in a given case. This contrasted with other keyword papers examined, where specific approaches or proposals were related to the key characteristic. Approximately one-third of the titles and abstracts reviewed in this set also included another trustworthy AI term, especially explainability, interpretability, and transparency, or the umbrella concept of XAI,

which echoes NIST’s statement that “accountability presupposes transparency” [1].

5 DISCUSSION

To help shape global norms for AI governance that will ultimately affect international commerce, diplomacy, and interoperability, stakeholders will need to monitor relevant national policy and scientific research. Examining the use of trustworthy AI terms in policy and research surfaces the frequency of their appearance, how international research efforts may align with high-level policy goals, and where a focus on research in and development of trustworthy AI is evident. Policymakers should remain aware of how the research community is using trustworthy AI terms so they can progress toward a more common understanding and track the development of commonly accepted techniques and frameworks, and vice versa. Defining and achieving consensus on trustworthy AI characteristics is a societal effort that must cross historic boundaries between the technical and policy communities. Our empirical findings highlight that overall there are discrepancies among the key terms use with regard to definitions in the policy and research communities, respectively, but the differences are not beyond alignment adjustments.

Within the research publication analysis we find a variance in term use to the policy documents with regard to frequency of mentions. Notably the terms accountability, transparency, fairness, and explainability are infrequently used in research publications’ titles and abstracts, compared to terms such as reliability, robustness, safety, security, and bias. The differences in term use is more nuanced than simply counting mentions, but this frequency comparison highlights the differences in the policy and research communities’ focus. Specifically, certain terms are more formally established research areas (e.g., XAI, security and privacy, and bias and fairness). Another aspect to policy and research term differences are the tensions between principles and expectations, particularly for terms such as transparency and privacy. 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. 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.

6 CONCLUSION

Through their published documents, the policy and research communities offer perspectives and approaches that will help society better avoid harm and achieve a positive impact with AI. Efforts to realize trustworthy AI are best helped, however, when these policymakers and researchers share a language—and an understanding—common to us all in our societal ambitions for trustworthy AI. Through comprehensive analysis of American, Australian, British, Canadian, and Japanese high-level policies and guidance on trustworthy AI, as well as the analysis of researcher use of key terms,

we form a better understanding of the challenges and opportunities ahead for the shared goal of trustworthy AI. We present our findings that policy documents generally agree on the importance of six concepts: accountability, explainability, fairness, privacy, security, and transparency, and that while these terms are echoed in the research literature, their definition can be different. Additionally, there are inconsistencies between policy and research term frequencies, highlighting the different focuses of each group on trustworthy AI. To communicate clearly and develop effective solutions, policymakers and researchers must share terminology and definitions to avoid talking past progress.

REFERENCES

- [1] NIST AI. 2023. Artificial Intelligence Risk Management Framework (AI RMF 1.0). (2023).
- [2] Sajid Ali, Tamer Abuhmed, Shaker El-Sappagh, Khan Muhammad, Jose M Alonso-Moral, Roberto Confalonieri, Riccardo Guidotti, Javier Del Ser, Natalia Díaz-Rodríguez, and Francisco Herrera. 2023. Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Information Fusion* 99 (2023), 101805.
- [3] IEEE Standards Association. 2017. Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems. https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf. Accessed: 2023-11-20.
- [4] Blair Attard-Frost, Ana Brandusescu, and Kelly Lyons. 2023. The Governance of Artificial Intelligence in Canada: Findings and Opportunities from a Review of 84 AI Governance Initiatives. Available at SSRN (2023).
- [5] Science Australia Department of Industry and Resources. 2023. Australia’s AI Ethics Principles. <https://legalinstruments.oecd.org/en/instruments/oecd-legal-0449>. (Canberra: Australian Government, Accessed March 17, 2023).
- [6] Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: Pretrained Language Model for Scientific Text. In *EMNLP*. arXiv:arXiv:1903.10676
- [7] A Feder Cooper, Emanuel Moss, Benjamin Laufer, and Helen Nissenbaum. 2022. Accountability in an algorithmic society: relationality, responsibility, and robustness in machine learning. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 864–876.
- [8] James Dunham, Jennifer Melot, and Dewey Murdick. 2020. Identifying the development and application of artificial intelligence in scientific text. *arXiv preprint arXiv:2002.07143* (2020).
- [9] Francesca Foffano, Teresa Scantamburlo, and Atia Cortés. 2023. Investing in AI for social good: an analysis of European national strategies. *AI & society* 38, 2 (2023), 479–500.
- [10] Japan Council for Social Principles of Human-centric AI. 2019. Social Principles of Human-Centric AI. <https://ai.bsa.org/wp-content/uploads/2019/09/humancentricai.pdf>. Accessed: 2023-11-20.
- [11] The White House. 2021. The United States, Joined by Allies and Partners, Attributes Malicious Cyber Activity and Irresponsible State Behavior to the People’s Republic of China. <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/>. Accessed: 2023-11-20.
- [12] Anna Jobin, Marcello Lenca, and Effy Vayena. 2019. The global landscape of AI ethics guidelines. *Nature machine intelligence* 1, 9 (2019), 389–399.
- [13] Joe King, Betty Zhang, Hanif Mahboobi, and Shantu Roy. 2021. Model Explainability with AWS Artificial Intelligence and Machine Learning Solutions. <https://docs.aws.amazon.com/whitepapers/latest/model-explainability-aws-ai-ml/interpretability-versus-explainability.html>. Accessed: 2023-11-20.
- [14] United Kingdom. 2018. Data Protection Act. <https://www.legislation.gov.uk/ukpga/2018/12/contents/enacted>. Accessed: 2023-11-20.
- [15] Johann Laux, Sandra Wachter, and Brent Mittelstadt. 2024. Trustworthy artificial intelligence and the European Union AI act: On the conflation of trustworthiness and acceptability of risk. *Regulation & Governance* 18, 1 (2024), 3–32.
- [16] Gurmeet Singh Manku, Arvind Jain, and Anish Das Sarma. 2007. Detecting near-duplicates for web crawling. In *Proceedings of the 16th international conference on World Wide Web*. 141–150.
- [17] OECD. 2023. Recommendation of the Council on OECD Legal Instruments Artificial Intelligence. <https://legalinstruments.oecd.org/en/instruments/oecd-legal-0449>. OECD/LEGAL/0449 Accessed: 2023-11-20.
- [18] The Government of Canada. 2019. Policy on Service and Digital. <https://www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=32603>. Accessed: 2023-11-20.
- [19] The Government of Canada. 2023. Directive on Automated Decision-Making. <https://www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=32592>. Accessed: 2023-11-20.

- [20] The Government of Canada. 2023. Privacy Act. <https://laws-lois.justice.gc.ca>. Accessed: 2023-11-20.
- [21] The Government of Canada. 2023. Responsible Use of Artificial Intelligence, Guiding Principles. <https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai.html>. Accessed: 2023-11-20.
- [22] Executive Office of the President. 2019. *E.O. 13859;84 FR 3967;2019-02544*. 3967–3972 pages.
- [23] Executive Office of the President. 2020. *E.O. 13960;85 FR 78939;2020-27065*. 78939–78943 pages.
- [24] United Kingdom Information Commissioner’s Office. 2018. UK GDPR guidance and resources. <https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/>. Accessed: 2023-11-20.
- [25] United Kingdom Information Commissioner’s Office. 2023. Guidance on AI and data protection. <https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/guidance-on-ai-and-data-protection/>. Accessed: 2023-11-20.
- [26] United Kingdom Information Commissioner’s Office and The Alan Turing Institute. 2022. Explaining decisions made with AI’s. <https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/explaining-decisions-made-with-artificial-intelligence/>. Accessed: 2023-11-20.
- [27] Japan Expert Group on How AI Principles Should be Implemented. 2022. Integrated Innovation Strategy Promotion Council, Governance Guidelines for Implementation of AI Principles. https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/pdf/20220128_2.pdf. Accessed: 2023-11-20.
- [28] International Standards Organization. 2022. Trustworthiness – Vocabulary. <https://www.iso.org/obp/ui/#iso:std:iso-iec:ts:5723:ed-1:v1:en>. Accessed: 2023-11-20.
- [29] Cecilia Panigutti, Ronan Hamon, Isabelle Hupont, David Fernandez Llorca, Delia Fano Yela, Henrik Junklewitz, Salvatore Scalzo, Gabriele Mazzini, Ignacio Sanchez, Josep Soler Garrido, et al. 2023. The role of explainable AI in the context of the AI Act. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 1139–1150.
- [30] Giada Pistilli, Carlos Muñoz Ferrandis, Yacine Jernite, and Margaret Mitchell. 2023. Stronger Together: on the Articulation of Ethical Charters, Legal Tools, and Technical Documentation in ML. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 343–354.
- [31] Charles Radclyffe, Mafalda Ribeiro, and Robert H Wortham. 2023. The assessment list for trustworthy artificial intelligence: A review and recommendations. *Frontiers in Artificial Intelligence* 6 (2023), 1020592.
- [32] Huw Roberts, Josh Cows, Jessica Morley, Mariarosaria Taddeo, Vincent Wang, and Luciano Floridi. 2021. The Chinese approach to artificial intelligence: an analysis of policy, ethics, and regulation. *AI & society* 36 (2021), 59–77.
- [33] Stephen Cory Robinson. 2020. Trust, transparency, and openness: How inclusion of cultural values shapes Nordic national public policy strategies for artificial intelligence (AI). *Technology in Society* 63 (2020), 101421.
- [34] Daniel Schiff, Justin Biddle, Jason Borenstein, and Kelly Laas. 2020. What’s next for ai ethics, policy, and governance? a global overview. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 153–158.
- [35] Giulia Vilone and Luca Longo. 2021. Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion* 76 (2021), 89–106.
- [36] T. Russell Vought. 2020. OMB Memorandum M-21-06: Guidance for Regulation of Artificial Intelligence Applications. <https://www.whitehouse.gov/wp-content/uploads/2020/11/M-21-06.pdf>. Accessed: 2023-11-20.

A DOCUMENT SELECTION CRITERION

Selecting the AI documents or policies most influential in a nation is challenging both because there has been a rapid proliferation of documents and because there is no clear, singular governing agency in any of these countries. We examined multiple documents and engaged with government experts to identify the optimal choices for document analysis for our research. For example, the UK has established law (i.e. the Human Rights Act and GDPR), proposed laws (i.e. the Data Protection and Digital Information Bill introduced in 2022), AI specific interpretations of laws (i.e. the ICO Guide to GDPR), and AI specific policy documents and statements (i.e. the National AI Strategy (published in 2021, updated in 2022), “Establishing a pro-innovation approach to regulating AI,” and the National Data Strategy). Reviewing these and other documents, we

applied our selection criteria to guide our final choices. Because the UK had so many documents, we took special note of how each discussed key principles, the references contained within each of these documents, and we also consulted UK government officials to ensure we properly understood which documents were most influential.

To select countries for our analysis, we went through a four-step filtering process. First, we examined statements by close U.S. allies so that our analysis might inform U.S. diplomatic efforts to establish trustworthy AI norms. Accordingly, we selected countries in the North Atlantic Treaty Organization, the Quadrilateral Security Dialogue, or the UK – United States of America Agreement (Five Eyes). Next, from this subset of countries, we reviewed the OECD AI Policy Observatory’s repository of documents to identify countries with AI policy documents that would meet our document inclusion criteria. Third, we eliminated any countries not listed as one of the top 15 highest investing countries in AI as identified by the Emerging Technology Observatory’s Country Activity Tracker. This third step eliminated all but nine countries: Australia, Canada, France, Germany, Japan, the Netherlands, the United Kingdom, and the United States. Given available resources and our interest in producing timely analysis for decision-makers, we chose to examine only 5 of these countries, specifically selecting ones that represented geographic diversity: Australia, Canada, Japan, and the United Kingdom, in addition to the United States. We chose the UK over France and Germany because of the UK’s greater investments and research in AI.

B POLICY DOCUMENT DESCRIPTIONS

Australia: In 2019 the Australian Department of Industry, Science, and Resources published a set of eight principles and corresponding descriptions to guide public and private use of AI [5]. This voluntary AI ethics framework is designed to ensure that AI use is safe, secure, and reliable by building public trust in AI products, increasing consumer loyalty in AI-powered services, and enabling Australians to benefit from AI. Of note, the document cites the Institute of Electrical and Electronics Engineers (IEEE) report, “Ethically Aligned Design,” as a source of inspiration and guidance [3].

Canada: With the goal to regulate the Canadian government’s use of any AI-enabled “system, tool, or statistical models used to recommend or make an administrative decision about a client”, Canada’s Directive on Automated Decision-Making was released in March 2019 and updated in 2021 and 2023 [19]. Building on prior Canadian legislation, including the Policy on Government and Digital [18] and the Privacy Act [20], the Directive includes impact assessment levels that guide the application of governance requirements to AI. In summary 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”, and *Level IV*) “Decisions will often lead to impacts that are irreversible and perpetual.” Additionally, the Directive on Automated Decision-Making is augmented by the Canadian government’s Responsible Use of Artificial Intelligence-Guiding Principles [21]. These AI guiding principles are directly aligned with Canada’s administrative law principles

and were designed alongside the Directive on Automated Decision-Making through facilitated workshops and published white papers, involving input from stakeholders in government, industry, and academia.

Japan: Japan's Society 5.0 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" [10]. The Social Principles of Human-Centric AI were designed as a part of Society 5.0 and are a voluntary set of AI guidelines for Japanese companies. Influenced by these social principles, an expert group on how AI principles should be implemented published Governance Guidelines for Implementation of AI Principles in 2022 [27]. Containing examples and target behavior for AI systems that could negatively impact society, the governance guidelines were constructed for Japanese companies to use as a reference point when developing governance mechanisms for AI, though they are not legally binding.

United Kingdom: The UK General Data Protection Regulation and Data Protection Act (UK GDPR) has been applied to AI systems and models, as most current AI technology relies heavily on data use [14]. For clarity on the UK GDPR, the Information Commissioner's Office (ICO) produced the "Guide to the UK General Data Protection Regulation", containing specific guidance on how the UK GDPR applies to AI systems [24]. We examined both the GDPR and its implementation guidance, we selected the "Guide to the UK GDPR" [25] instead of the GDPR itself, as the former

is organized for practitioners working with AI systems, and thus more consolidated and specific to AI. 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.

United States: Executive Order 13960, Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government, was issued in 2020 and was the current Executive Order at the time of analysis [23]. 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 directs 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" [22, 23]. Executive Order 13859 (published in 2019), Maintaining American Leadership in Artificial Intelligence, provides guidance to executive agencies on how to support the research and development of AI-enabled systems, but only mentions ethical AI terms in passing [22]. The Office of Management and Budget published amplifying guidance on EO13859, The Guidance for Regulation of Artificial Intelligence Applications, which specifically addresses key trustworthy AI terms [36]. Thus, we include EO13960 and the OMB guidance, but not EO13859, in our analysis.