

The Evolution of AI Governance

Simon Chesterman , National University of Singapore and AI Singapore

Yuting Gao , ESCP Business School

Jungpil Hahn , National University of Singapore and AI Singapore

Valerie Sticher , University of Zurich

Many companies and a growing number of governments have guides, frameworks, or principles on their use of artificial intelligence (AI). Analyzing the evolving practice of releasing such documents and the language they use offers important insights into how norms around AI are spreading and changing— and where they might go next.

Governance of new technologies often vacillates between irrelevance and excess. Regulating too little or too late makes governance an afterthought; overregulation risks constraining innovation or driving it elsewhere.⁸ It is

striking that the development of artificial intelligence (AI) for most of its history saw few serious discussions of governance at all. Indeed, as recently as 2011, there were serious calls for limited immunity to be granted to developers, lest the possibility of lawsuits deter investment.⁶

Data protection rules, especially in Europe, were applied to certain applications of AI. However, for half a century the state of the art went little further than

Digital Object Identifier 10.1109/MC.2024.3381215
Date of current version: 27 August 2024

science fiction author Isaac Asimov's three laws of robotics.⁵

Those laws (do not harm humans, obey orders, and protect yourself) are regularly cited in the literature on regulating new technology. Like the Turing test, however, they are more of a cultural touchstone than serious scientific or regulatory proposal. The laws assume, for example, that the target of regulation is physically embodied robots with human-level intelligence, an example of the android fallacy. Another critique is that they place obligations on the technology itself, rather than the people creating it.⁴ In fact, Asimov's laws were not "laws" in the sense of commands to be enforced by the state. They were, rather, encoded into the positronic brains of his fictional creations: constraining what robots *could* do, rather than specifying what they *should*.

Of most direct concern for this article, the idea that relevant ethical principles could be reduced to a few dozen words, or that those words might be encoded in a manner interpretable by an AI system, oversimplifies the nature of ethics and of law. Nonetheless, in 2007 it was reported that Korea had considered using them as the basis for a proposed Robot Ethics Charter.⁷ That was one of many attempts to codify norms governing "robots" since the turn of the century, accelerating in the wake of the First International Symposium on Roboethics in Sanremo, Italy, in 2004. The European Robotics Research Network produced its "Roboethics Roadmap" in 2006, while the first multidisciplinary set of principles for robotics was adopted at a Robotics Retreat held by two British Research Councils in 2010.

Much has happened since. The years after 2016 saw a proliferation of guides, frameworks, and principles

focused specifically on AI. Some were the product of conferences or industry associations, notably the Partnership on AI's Tenets (2016), the Future of Life Institute's Asilomar AI Principles (2017), the Beijing Academy of Artificial Intelligence's Beijing AI Principles (2019), and IEEE's Ethically Aligned Design (2019). Others were drafted by individual companies, including Microsoft's Responsible AI Principles, IBM's Principles for Trust and Transparency, and Google's AI Principles, all published in the first half of 2018. This proliferation was driven by the machine learning revolution of the 2010s, alongside the Cambridge Analytica scandal of 2016. The former demonstrated the potential of AI; the latter showed that potential harms went far beyond biased decisions or errant autonomous vehicles to include impacting elections in the most powerful country on the planet.^{7,10,12} Fears of more serious political interference, particularly disinformation campaigns using generative AI, were widely debated in 2024.

In the intervening years, analysis of when, how, and why to govern AI systems became more sophisticated, drawing on regulatory theory as well as more nuanced understanding of what machine learning models can (and cannot) do.¹ Key questions include the scope of application (such as how "AI" is defined), the precise harms to be avoided, and how any governance regime should balance the opportunities afforded by AI against the risks.^{2,11}

Today, an ever-expanding number of companies and countries have adopted documents setting out principles to govern the development and use of AI. Early movers may have been technology companies in Western countries, but this has quickly become a global phenomenon. As the practice

of adopting such documents spreads, by sector and by country, overlapping language has converged on certain key principles with remarkable consistency. At the same time, there are important divergences in what those principles are understood to mean.^{18,19} This article analyzes the evolution of these AI governance principles, focusing on the spread of similar terms around the world and the differing ways in which those terms are understood, in theory as well as in practice. Although the field is rapidly developing, an analysis of the first eight years of serious AI governance debates provides valuable insights into the development, spread, and future directions of AI norms.

THE RISE OF AI GOVERNANCE

To gather representative data, we examined a wide range of documents issued by governments and companies.^a For governments, we searched for documents issued by all 193 member states of the United Nations. For companies, our focus was on the world's 100 largest companies by market capitalization.^{9,b} This selection strategy allowed us to capture not only AI developers but also its users. The full list of companies analyzed is included in [Table 1](#).

For all these cases, we conducted a systematic search on Google^c ([Table 2](#))

^aWhile our data are confined to documents issued by governments and companies, comparing them with policies established by other entities, such as ongoing multinational regulatory efforts, offers a promising avenue for future research.

^bFocusing on companies with a large market capitalization increases the likelihood of capturing the emergence of AI principles as smaller companies often lag in issuing AI governance documents.

^cThe following Google search terms were used: (AI OR artificial intelligence) AND (principles OR guidelines OR recommendations) AND ("[country/company name]"). The first thirty and fifteen search results for countries and companies, respectively, were screened. This procedure was developed iteratively after several rounds of pilot searches.

CERTIFYING AND REGULATING AI/ML-CENTRIC APPLICATIONS

TABLE 1. A list of companies, compiled May 2022 (top 100 by market capitalization based on <https://companiesmarketcap.com>, in alphabetical order).

Company	Sector	Company	Sector	Company	Sector
Abbott Laboratories	Health care	CVS Health	Health care	PetroChina	Others
AbbVie	Health care	Danaher	Health care	Pfizer	Health care
Accenture	Others	Eli Lilly	Health care	Philip Morris	Others
Adobe	Technology	Exxon Mobil	Others	Procter & Gamble	Others
Agricultural Bank of China	Financials	Hermés	Others	Prosus	Technology
Alibaba	Others	Home Depot	Others	QUALCOMM	Technology
Alphabet (Google)	Technology	Honeywell	Others	Raytheon Technologies	Technology
Amazon	Others	HSBC	Financials	Reliance Industries	Others
AMD	Technology	ICBC	Financials	Roche	Health care
American Express	Financials	Intel	Technology	Royal Bank Of Canada	Financials
Amgen	Health care	Johnson & Johnson	Health care	Salesforce	Technology
Anthem	Health care	JPMorgan Chase	Financials	Samsung	Technology
Apple	Technology	Kweichow Moutai	Others	Sanofi	Health care
ASML	Technology	L'Oréal	Others	Saudi Aramco	Others
AstraZeneca	Health care	Linde	Others	Shell	Others
AT&T	Financials	Lowe's Companies	Others	T-Mobile US	Others
Bank of America	Financials	LVMH	Others	Tata Consultancy Services	Technology
Bank of China	Financials	Mastercard	Others	Tencent	Technology
Berkshire Hathaway	Financials	McDonald	Others	Tesla	Others
BHP Group	Others	Medtronic	Health care	Texas Instruments	Technology
Bristol Myers Squibb	Health care	Meituan	Technology	Thermo Fisher Scientific	Health care
Broadcom	Technology	Merck	Health care	Toronto Dominion Bank	Financials
CATL	Others	Meta (Facebook)	Technology	TotalEnergies	Others
Charles Schwab	Financials	Microsoft	Technology	Toyota	Others
Chevron	Others	Morgan Stanley	Financials	TSMC	Technology
China Construction Bank	Financials	Nestle	Others	Union Pacific Corporation	Others
China Mobile	Others	Nextera Energy	Others	United Parcel Service	Others
Cisco	Others	Nike	Others	UnitedHealth	Health care
CM Bank	Financials	Novartis	Health care	Verizon	Others
Coca-Cola	Others	Novo Nordisk	Health care	Visa	Others
Comcast	Others	NVIDIA	Technology	Walmart	Others
Commonwealth Bank	Financials	Oracle	Technology	Walt Disney	Others
ConocoPhillips	Others	Pepsico	Others	Wells Fargo	Financials
Costco	Others				

TABLE 2. A list of documents coded by (a) entity type “country” or (b) entity type “company” (in alphabetical order by type). All coded data can be accessed at <https://aisingapore.org/governance/resources>.

(a) Name of the document (official English translation in parentheses if the original is not in English)	Year	Entity name	Entity type	Region if a country
Australia’s Artificial Intelligence Ethics Framework	2019	Australia	Country	Western Europe and others
Australia’s Artificial Intelligence Action Plan	2021	Australia	Country	Western Europe and others
AI4Belgium Report	2019	Belgium	Country	Western Europe and others
Concept for the Development of Artificial Intelligence in Bulgaria Until 2030	2020	Bulgaria	Country	Eastern Europe
Responsible Use of Artificial Intelligence (AI)	2019	Canada	Country	Western Europe and others
Directive on Automated Decision-Making	2021	Canada	Country	Western Europe and others
新一代人工智能发展规划的通知 (New Generation Artificial Intelligence Development Plan)	2017	China	Country	Asia-Pacific
发展负责任的人工智能：新一代人工智能治理原则 (transl.: New Generation of Artificial Intelligence: Developing Responsible Artificial Intelligence)	2019	China	Country	Asia-Pacific
新一代人工智能伦理规范 (transl.: Ethical Norms for New Generation Artificial Intelligence)	2021	China	Country	Asia-Pacific
Ethical Framework for Artificial Intelligence in Colombia	2020	Colombia	Country	Latin America and Caribbean
National Artificial Intelligence Strategy of the Czech Republic	2019	Czech Republic	Country	Eastern Europe
National Strategy for Artificial Intelligence	2019	Denmark	Country	Western Europe and others
Egypt National Strategy for Artificial Intelligence	2021	Egypt	Country	Africa Group
Estonia’s National Artificial Intelligence Strategy 2019–2021	2019	Estonia	Country	Eastern Europe
Finland’s Age of Artificial Intelligence	2017	Finland	Country	Western Europe and others
Leading the Way Into the Age of Artificial Intelligence	2019	Finland	Country	Western Europe and others
How Can Humans Keep the Upper Hand? Report on the Ethical Matters Raised by Algorithms and Artificial Intelligence	2017	France	Country	Western Europe and others
AI for Humanity	2018	France	Country	Western Europe and others
Automated and Connected Driving	2017	Germany	Country	Western Europe and others
Artificial Intelligence Strategy of the German Federal Government (2020 Update)	2020	Germany	Country	Western Europe and others
Democratising AI: A National Strategy for Greece	2020	Greece	Country	Western Europe and others
Hungary’s Artificial Intelligence Strategy	2020	Hungary	Country	Eastern Europe
Tamil Nadu: Safe & Ethical Artificial Intelligence Policy 2020	2020	India	Country	Asia-Pacific
Responsible AI—#AIforAll: Approach Document for India (Part 1—Principles for Responsible AI)	2021	India	Country	Asia-Pacific

(Continued)

CERTIFYING AND REGULATING AI/ML-CENTRIC APPLICATIONS

TABLE 2. (Continued.) A list of documents coded by (a) entity type “country” or (b) entity type “company” (in alphabetical order by type). All coded data can be accessed at <https://aisingapore.org/governance/resources>.

(a) Name of the document (official English translation in parentheses if the original is not in English)	Year	Entity name	Entity type	Region if a country
Responsible AI—#AIforAll: Approach Document for India (Part 2—Operationalizing Principles for Responsible AI)	2021	India	Country	Asia-Pacific
AI—Here for Good: A National Artificial Intelligence Strategy for Ireland	2021	Ireland	Country	Western Europe and others
White Paper on Artificial Intelligence at the Service of Citizens	2018	Italy	Country	Western Europe and others
Social Principles of Human-Centric AI	2019	Japan	Country	Asia-Pacific
AI Strategy 2019	2019	Japan	Country	Asia-Pacific
Governance Guidelines for Implementation of AI Principles	2021	Japan	Country	Asia-Pacific
Advisory Report on Development of an Artificial Intelligence Strategy for Lebanon	2020	Lebanon	Country	Asia-Pacific
Lithuanian Artificial Intelligence Strategy	2019	Lithuania	Country	Eastern Europe
Artificial Intelligence: A Strategic Vision for Luxembourg	2019	Luxembourg	Country	Western Europe and others
Towards Trustworthy AI: Malta Ethical AI Framework for Public Consultation	2019	Malta	Country	Western Europe and others
Malta’s National AI Strategy	2019	Malta	Country	Western Europe and others
Mauritius Artificial Intelligence Strategy	2018	Mauritius	Country	Africa Group
Toward an AI Strategy in Mexico	2018	Mexico	Country	Latin America and Caribbean
Artificial Intelligence in Mexico: A National Agenda	2020	Mexico	Country	Latin America and Caribbean
Strategic Action Plan for AI	2019	Netherlands	Country	Western Europe and others
AI: Shaping a Future New Zealand	2018	New Zealand	Country	Western Europe and others
National Strategy for Artificial Intelligence	2020	Norway	Country	Western Europe and others
National Artificial Intelligence Strategy: First Draft of Peruvian National AI Strategy	2021	Peru	Country	Latin America and Caribbean
National AI Strategy Roadmap for the Philippines	2021	Philippines	Country	Asia-Pacific
AI Portugal 2030	2019	Portugal	Country	Western Europe and others
Blueprint: National Artificial Intelligence Strategy for Qatar	2019	Qatar	Country	Asia-Pacific
National Strategy for Artificial Intelligence 2019	2019	Republic of Korea	Country	Asia-Pacific
Romania in the Era of Artificial Intelligence: A Strategy for the Development and Adoption of AI Technology at a Country Level	2019	Romania	Country	Eastern Europe

(Continued)

TABLE 2. (Continued.) A list of documents coded by (a) entity type “country” or (b) entity type “company” (in alphabetical order by type). All coded data can be accessed at <https://aisingapore.org/governance/resources>.

(a) Name of the document (official English translation in parentheses if the original is not in English)	Year	Entity name	Entity type	Region if a country
Decree of the President of the Russian Federation on the Development of Artificial Intelligence in the Russian Federation	2019	Russian Federation	Country	Eastern Europe
Development of Artificial Intelligence	2019	Russian Federation	Country	Eastern Europe
Realizing Our Best Tomorrow: Strategy Narrative	2020	Saudi Arabia	Country	Asia-Pacific
Strategy for the Development of Artificial Intelligence in the Republic of Serbia for the Period 2020–2025	2019	Serbia	Country	Eastern Europe
Principles to Promote Fairness, Ethics, Accountability and Transparency (FEAT) in the Use of Artificial Intelligence and Data Analytics in Singapore’s Financial Sector	2018	Singapore	Country	Asia-Pacific
National Artificial Intelligence Strategy	2019	Singapore	Country	Asia-Pacific
Model AI Governance Framework (2nd ed.)	2020	Singapore	Country	Asia-Pacific
Artificial Intelligence in Healthcare Guidelines (AIHGle)	2021	Singapore	Country	Asia-Pacific
Action Plan for the Digital Transformation of Slovakia for 2019–2022	2019	Slovakia	Country	Eastern Europe
Guidelines on Artificial Intelligence for the Confederation: General Frame of Reference on the Use of Artificial Intelligence Within the Federal Administration	2020	Switzerland	Country	Western Europe and others
National Approach to Artificial Intelligence	2018	Sweden	Country	Western Europe and others
UAE Strategy for Artificial Intelligence 2031	2018	United Arab Emirates	Country	Asia-Pacific
Artificial Intelligence Principles & Ethics	2019	United Arab Emirates	Country	Asia-Pacific
Artificial Intelligence Sector Deal	2018	United Kingdom	Country	Western Europe and others
AI in the UK: Ready, Willing and Able?	2018	United Kingdom	Country	Western Europe and others
National AI Strategy	2021	United Kingdom	Country	Western Europe and others
Preparing for the Future of Artificial Intelligence	2016	United States	Country	Western Europe and others
Artificial Intelligence, Automation, and the Economy	2016	United States	Country	Western Europe and others
National Artificial Intelligence Research and Development Strategic Plan: 2019 Update	2019	United States	Country	Western Europe and others
Principles of Artificial Intelligence Ethics for the Intelligence Community	2020	United States	Country	Western Europe and others
Artificial Intelligence Strategy for the Digital Government	2019	Uruguay	Country	Latin America and Caribbean

(Continued)

CERTIFYING AND REGULATING AI/ML-CENTRIC APPLICATIONS

TABLE 2. (Continued.) A list of documents coded by (a) entity type “country” or (b) entity type “company” (in alphabetical order by type). All coded data can be accessed at <https://aisingapore.org/governance/resources>.

(b) Name of the document (official English translation in parentheses if the original is not in English)	Year	Entity name	Entity type	Sector if a company
Responsible AI and Robotics: An Ethical Framework	2019	Accenture	Company	Industrials
Responsible AI: From Principles to Practice	2021	Accenture	Company	Industrials
Adobe’s Commitment to AI Ethics	2021	Adobe	Company	Technology
Artificial Intelligence at Google: Our Principles	2021	Alphabet (Google)	Company	Technology
AstraZeneca Data and AI Ethics	2020	Astrazeneca	Company	Health care
Artificial Intelligence at AT&T: Our Guiding Principles	2019	AT&T	Company	Financials
AI Ethics Case Study: Commonwealth Bank of Australia	2021	Commonwealth Bank	Company	Financials
HSBC’s Principles for the Ethical Use of Data and AI	2020	HSBC	Company	Financials
Artificial Intelligence: The Public Policy Opportunity	2017	Intel	Company	Technology
Intel’s AI Privacy Policy White Paper: Protecting Individuals’ Privacy and Data in the Artificial Intelligence World	2018	Intel	Company	Technology
Intel’s Recommendations for the U.S. National Strategy on Artificial Intelligence	2019	Intel	Company	Technology
Thoughtful, Strategic Implementation of AI Builds Trust: The Five Pillars of Mastercard AI	2020	Mastercard	Company	Industrials
Facebook’s Five Pillars of Responsible AI	2021	Meta (Facebook)	Company	Technology
Responsible AI	2018	Microsoft	Company	Technology
Novartis’ Commitment to the Ethical and Responsible Use of Artificial Intelligence (AI) Systems	2021	Novartis	Company	Health care
Sustainability Review	2021	Prosus	Company	Technology
Ensuring AI Remains a Force for Good	2021	Royal Bank Of Canada	Company	Financials
Borealis AI Launches RESPECT AI Program to Bring Ethical and Responsible AI to All	2020	Royal Bank Of Canada	Company	Financials
AI Ethics Maturity Model	2021	Salesforce	Company	Technology
Artificial Intelligence	2021	Samsung	Company	Technology
Guiding Principles for AI at Sanofi	2021	Sanofi	Company	Health care
Ethical AI: A New Strategic Imperative for Recruiting and Staffing	2021	Tata Consultancy Services	Company	Technology
Ensuring Safe, Secure, and Sustainable AI	2021	Tata Consultancy Services	Company	Technology
“ARCC”: An Ethical Framework for Artificial Intelligence	2020	Tencent	Company	Technology

to determine whether a government [Table 2(a)] or a company [Table 2(b)] had released documents containing AI governance principles. In some cases, the top search results led directly to documents containing AI governance principles; in others, they contained a link or a simple reference to such a document, in which case we extended the search with the information provided. If none of the top results contained such a reference, we assumed that no official document exists (although there may be internal documents with similar purposes). We also referred to recent scholarship examining national AI strategies issued by governments worldwide, augmenting our dataset.¹⁵

Our data show that the introduction of AI governance principles is a relatively recent phenomenon: in 2016, no company and only one country, the United States, had released official documents explicitly governing AI. However, the number quickly increased in the years that followed. Generally, Western Europe and other countries were early adopters in defining relevant AI governance principles, with the rest of the world following suit (Figure 1). Among companies, unsurprisingly, principles first emerged in the tech sector, with other sectors catching up from 2018 onward (Figure 2).

Despite this clear trend, many countries and companies still lack official documents outlining AI governance principles. By the start of 2022, fewer than one in three governments (60 out of 193 states) and only one in five of the largest 100 companies (21 out of 100 companies) had released documents governing their use of AI systems, a proportion likely to be considerably lower in smaller companies. Even in the technology sector, fewer than half of the largest technology

companies (10 out of 21) had released an AI governance document by 2022, with the ratio much lower in sectors like finance (23.53% = 4/17) or health care (15.79% = 3/19), where AI plays an increasingly important role.

As AI use becomes more widespread, demands for clearer policies on its governance are increasing. An analogy may be drawn with the spread of personal data laws and policies, with most jurisdictions

now having data protection laws and more and more organizations adopting privacy policies that go beyond mere compliance with those laws.

The difference is that data protection and privacy embrace a relatively clear set of activities and concerns: data about identifiable individuals and potential harms arising from their collection, use, and disclosure. AI refers to a far wider range of technologies

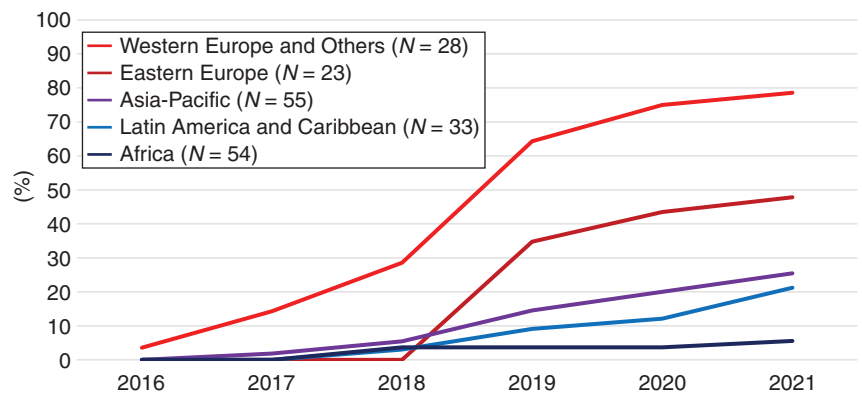


FIGURE 1. The percentage of countries in a region²⁰ that have released at least one AI governance principles document. The total number of countries in each region is in parentheses.

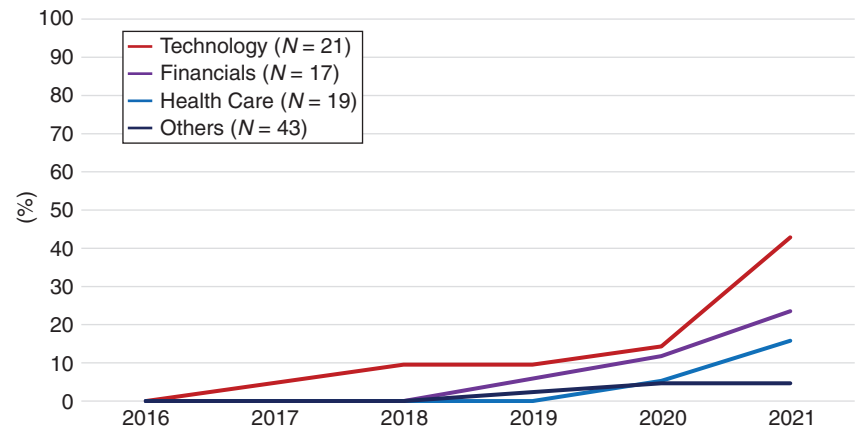


FIGURE 2. The percentage of companies in a sector that have released at least one AI governance principles document. The total number of companies in each sector is in parentheses.

and techniques, with similarly heterogeneous concerns about possible negative consequences.³

That diversity of use cases and harms is reflected in the more detailed examination of AI governance policies that follows. For, despite significant convergence in theory concerning the applicable norms, there is considerable disagreement about what those norms mean in practice.

THAT DIVERSITY OF USE CASES
AND HARMS IS REFLECTED IN THE
MORE DETAILED EXAMINATION OF AI
GOVERNANCE POLICIES THAT FOLLOWS.

A TAXONOMY OF AI PRINCIPLES

To capture the content of AI governance policies, we engaged in an iterative process to create a taxonomy of AI principles. We used existing literature as a starting point, building in particular on an early mapping of AI ethics guidelines.¹³ We then analyzed government and company documents to refine and specify the principles that are currently in use, adding and merging similar concepts to best reflect the current landscape of AI principles. One challenge we faced was that the examined documents vary greatly in formality and format, ranging from fully fledged AI policies to mere descriptions of AI principles on a company's website.^d However, while the depth of terminology naturally varied across these documents, remarkable

consistency emerged in the basic choice of terms they employed. This iterative process led us to coalesce around six primary AI principles.

1. *Fairness*: The fairness principle typically addresses aspects of bias arising from AI systems and the inclusiveness of these systems. This principle can encompass up to three aspects.

Nondiscrimination ensures that the design, development, and deployment of AI systems does not introduce unfair biases or exacerbate existing ones. *Inclusiveness in impact* emphasizes the need for AI development and deployment to benefit society at large, rather than a narrow subset of individuals. *Inclusiveness in design* involves incorporating individuals from diverse backgrounds into design and development processes. Each of these three aspects contributes individually or collectively to the broader concept of fairness in AI.

2. *Accountability*: The accountability principle is concerned with the regulation of AI systems through legal governance, risk management, and compliance mechanisms. This

principle seeks to minimize the risks associated with AI and provide appropriate remedies for losses.

3. *Transparency/explainability*: The transparency/explainability principle is generally concerned with facilitating human understanding of AI systems, including their algorithms and decisions. The principle can be divided into three components, which contribute to the overarching goal of transparency and explainability in AI. The first component, *transparency*, involves designing, developing, and deploying AI systems that facilitate human oversight through openness to external scrutiny. *Explainability* ensures that AI systems are accessible and interpretable to laypersons, enabling individuals to understand the general workings of an AI system and the decision-making process behind specific outcomes. *Openness* encourages the open sharing of data, as well as promoting open source research and collaboration in the design and development of AI systems.
4. *Ethics/human-centricity*: The ethics/human-centricity principle focuses on aligning AI systems with human values, encompassing up to three specific aspects. *Benefiting humans* seeks to ensure that AI systems are designed, developed, and deployed for the good of humankind, advocating human-aligned and nonmaleficent AI. *Human control of technology*

^dTo avoid ambiguity, we included documents irrespective of formality, encompassing both general statements of principles and more formal commitments to adherence.

emphasizes the necessity for humans to remain in control and have effective oversight of AI systems. Last, *human rights compliance* aims to ensure that the design, development, and deployment of AI complies with international human rights frameworks and norms.

5. *Safety/security*: The safety/security principle addresses the reliable operation of AI systems and their vulnerability to external threats. The *safety* dimension seeks to ensure that AI systems operate reliably without causing harm to living beings. It emphasizes the importance of guarding against possible misuse of AI systems. The *security* dimension focuses on adequately protecting AI systems from external threats, such as cyberattacks.
6. *Privacy*: Privacy is concerned with ensuring that data are appropriately handled in the design, development, and deployment of AI systems. This principle emphasizes user control in relation to data and compliance with existing privacy frameworks and norms.

In addition to the six primary principles, our analysis revealed other principles that were common but used less frequently. One such principle is *sustainability*, which concerns the long-term impact of AI development and deployment on humankind and the environment. This includes the impact of AI systems on future generations and any potential extinction risk posed by AI. *Professional responsibility of developers*, another principle that surfaced in several documents, points

to ethical obligations of individuals involved in the design, development, and deployment of AI systems. It emphasizes the role of developers in building safe and beneficial AI. Last, some documents highlighted the need for *technical competence*, pertaining to

analyzed for the presence of specific principles.^{g,h} As discussed in our taxonomy, some of the principles can be broken down into several components. We coded both at the level of components and at the level of principles (marked as referenced if it men-

HOWEVER, WHILE THE DEPTH OF TERMINOLOGY NATURALLY VARIED ACROSS THESE DOCUMENTS, REMARKABLE CONSISTENCY EMERGED IN THE BASIC CHOICE OF TERMS THEY EMPLOYED.

the technical expertise of individuals involved in the design, development, and deployment of AI systems.

CONVERGENCE BUT IN NAME ONLY?

In the development of the above taxonomy, we observed a remarkable consistency in the terminology used by governments and companies. To quantify this trend, we examined all government and company documents that were available in English.^e Our objective was to track the emergence of principles across various regions and sectors, rather than to offer a representative overview of the principles specific to each region and sector.^f

We organized all documents in a database, where each document was

tioned at least one of its components). For example, a document that contains a reference to nondiscrimination is coded as mentioning the fairness principle, even if it does not reference inclusiveness in impact or design.

Our data confirm the convergence in terminology over time at the level of principle. From the first releases to 2021, documents became more similar in terms of the principles they included. By 2021, more than 50% of documents included all six key principles that we coded (Figure 3). While principles like privacy and ethics/human-centricity were less often included in the earlier years, they

^gThe list of documents analyzed is included in Table 2. All coded data can be accessed at <https://aisingapore.org/governance/resources>.

^hThree research assistants participated in the coding process. To ensure consistent coding, we furnished them with detailed guidelines and examples at the project's outset and conducted regular meetings to discuss any instances of ambiguity. We further corroborated our findings with recent scholarship, notably Papyshv and Yarime,¹⁵ wherever information was accessible.

^eWe coded both original documents in English and official translations, including by reputable third-party institutions.

^fAchieving the latter is not feasible, as our coding is limited to existing documents. This limitation results in certain regions and sectors being disproportionately represented in our dataset (see Figure 1).

became more common from 2018 onward, as official documents on AI governance spread from the West to the rest of the world and from tech companies to other sectors.

A more detailed comparison between governments and companies shows that company documents tend to be more comprehensive than

government documents, although the gap is closing. This pattern is evident in five of the six primary principles: accountability, security, fairness, transparency/explainability, and ethics/human-centricity. Only in the concern for privacy do we see equal development between countries and companies. However, government documents more

frequently refer to other principles such as sustainability, professional responsibility of developers, and technical competence, with companies also gradually closing this gap.

Convergence in form may not, however, mean convergence in substance. What do different governments and companies actually mean when they refer to a specific principle? As we elaborate above, several principles can refer to diverse aspects of the design, development, and deployment of AI. For example, what we capture under the broad umbrella term “fairness” may refer to nondiscrimination, inclusiveness in impact, or inclusiveness in design. Our analysis shows that government and company actors are all more likely to mean nondiscrimination when including “fairness” as a principle (Figure 4). There are no significant differences in how governments and companies address the three aspects of fairness. This is supported by the *p* values obtained from chi-square tests for “nondiscrimination,” “inclusiveness in impact,” and

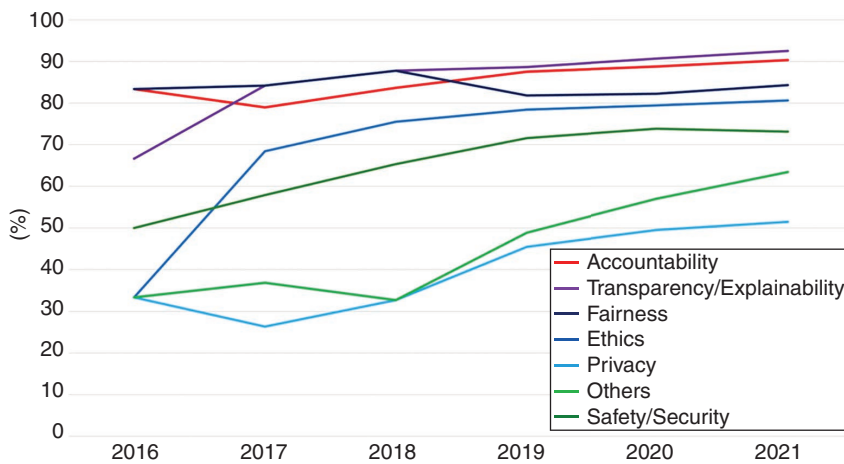


FIGURE 3. The percentage of AI governance principles documents that refer to a specific principle (countries and companies together).

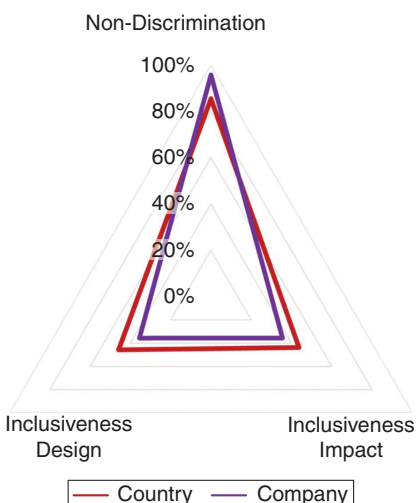


FIGURE 4. Diverse aspects of fairness (government versus company documents).

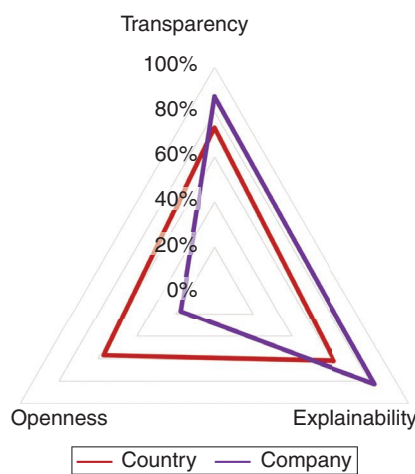


FIGURE 5. Diverse aspects of transparency/explainability (government versus company documents).

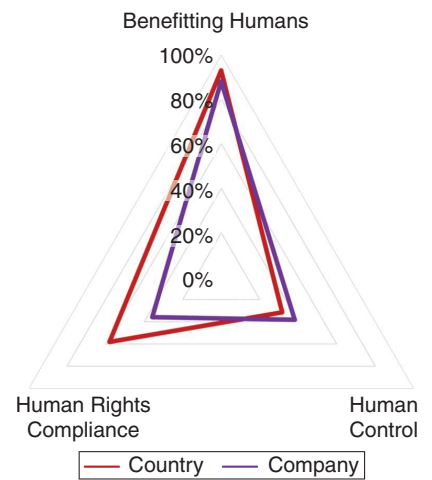


FIGURE 6. Diverse aspects of ethics/human-centricity (government versus company documents).

“inclusiveness in design,” which are 0.316, 0.746, and 0.654, respectively.

In other areas, there is divergence between governments and companies. In the area of transparency/explainability, for example, many government documents highlight “openness” as an aspect of transparency and explainability, whereas companies are less likely to include it within their principles (Figure 5).ⁱ Similarly, governments are more likely than companies to link compliance with human rights obligations under the auspices of ethics/human-centricity (Figure 6), although the difference is not statistically significant.^j

SURVIVAL OF THE FITTEST

As AI systems play ever greater roles across society and the economy, affecting the way we live, work, and play, countries and companies will see greater demand for accountability and appropriate levels of transparency in how those systems are deployed. Those demands are certain to grow as generative AI models play a greater role in the wider culture, as seen in the popularity of image and text generators such as Stable Diffusion and ChatGPT and the controversies surrounding their early (some would say premature) deployment.

AI governance policies have emerged as a means of showing citizens and consumers that these concerns are acknowledged. For some governments and companies, they have also offered

useful guardrails to limit certain potentially harmful deployments of AI or raise the costs for new entrants into the market. All of this is separate, however, from the impact that such guides, frameworks, and principles have in practice.¹⁴ Despite the fact that many such principles began in Western countries or technology companies, their spread around the world and across sectors has seen countervailing trends: convergence around similar language but with diverse interpretations of what that language means.

As the first efforts to move from governance by principle to regulation by law begin, most prominently in the European Union with its draft AI Act but also in the United States with President Biden’s Executive Order of 2023 and ongoing efforts in the United Nations in the lead up to the September 2024 Summit of the Future, governments will need to decide whether, when, and how to operationalize these norms. At the same time, some companies may find that there is a market value to embracing “responsible AI,”

ABOUT THE AUTHORS

SIMON CHESTERMAN is a professor in the Faculty of Law at the National University of Singapore, 259776, Republic of Singapore, and senior director of AI governance at AI Singapore, 117602, Republic of Singapore. His research interests include regulation and governance of AI. Chesterman received a D.Phil. in law from the University of Oxford. Contact him at chesterman@nus.edu.sg.

YUTING GAO is an assistant professor in the Department of Information & Operations Management at ESCP Business School, 28035 Madrid, Spain. Her research interests include human–computer interaction, human–AI interaction, and the economic and social impacts of the Internet. Gao received a Ph.D. in information systems and analytics from the National University of Singapore. Contact her at ygao@escp.eu.

JUNGPIL HAHN is a professor at the School of Computing of National University of Singapore, 119391, Republic of Singapore, and deputy director of AI governance at AI Singapore, 117602, Republic of Singapore. His research interests include open innovation, organizational learning and knowledge management, and human–computer interaction. Hahn received a Ph.D. in information and decision sciences from the University of Minnesota. Contact him at jungpil@nus.edu.sg.


VALERIE STICHER is a postdoctoral researcher in the Department of Political Science at the University of Zurich, 8050 Zurich, Switzerland. Her research interests include the impacts of AI on strategic decision-making in armed conflicts. Sticher received a Ph.D. in governance and global affairs from the University of Leiden. Contact her at sticherv@gmail.com.

ⁱChi-square tests show that the differences in “transparency” and “explainability” are not statistically significant, while the difference in openness is statistically significant. The *p* values for “transparency,” “explainability,” and “openness” are 0.263, 0.092, and 0.001, respectively.

^jChi-square tests show that the differences in all three aspects are not statistically significant. The *p* values for “benefiting humans,” “human control technology,” and “human rights compliance” are 0.727, 0.835, and 0.150, respectively.

above and beyond compliance or minimization of legal exposure. Against this, however, the rush to deploy large-language models, with accounts of downsizing or marginalization of ethics and responsible AI teams, points to the commercial pressures against robust governance.

The flaw in many of these debates has long been the assumption that an updated version of Asimov's laws would "solve" the problem of AI risk.¹⁶ The problem is that Asimov was a much better author than legislator; indeed, if his laws had worked, his literary career would have been short. In fact, even in the first story where he introduced the laws, they failed.

Having reached a broad consensus on the theoretical principles guiding AI design, development, and deployment, the challenge now is to apply these principles to specific use cases and illustrate their practical implications. This calls for future research that compares stated principles with actions taken to adhere to them. A deeper understanding of the institutionalization of governance and regulatory frameworks, and how they translate into action, is crucial to ensure that the terminology now frequently and consistently employed by governments and corporate actors is effectively put into action.^{2,11} 

ACKNOWLEDGMENT

We thank Rishika Madan, Khin Yadnanar Oo, Suzuki Tomoe, Yu Shi Jie, and Yonggang Li for invaluable research assistance in gathering and analyzing the data. We also thank several anonymous reviewers for their suggestions on improvements to the text. This article has supplementary downloadable material available at <https://doi.org/>

10.1109/MC.2024.3381215, provided by the authors.

REFERENCES

1. C. Alfonsi, "Taming tech giants requires fixing the revolving door," *Kennedy School Rev.*, vol. 19, pp. 166–170, 2019. <https://www.proquest.com/docview/2316725199/fulltextPDF/4F677B3FE2CA46B1PQ/32>
2. V. Almeida, L. S. Mendes, and D. Doneda, "On the development of AI governance frameworks," *IEEE Internet Comput.*, vol. 27, no. 1, pp. 70–74, Jan./Feb. 2023, doi: [10.1109/MIC.2022.3186030](https://doi.org/10.1109/MIC.2022.3186030).
3. G. Auld, A. Casovan, A. Clarke, and B. Faveri, "Governing AI through ethical standards: Learning from the experiences of other private governance initiatives," *J. Eur. Public Policy*, vol. 29, no. 11, pp. 1822–1844, 2022, doi: [10.1080/13501763.2022.2099449](https://doi.org/10.1080/13501763.2022.2099449).
4. J. M. Balkin, "The three laws of robotics in the age of big data," *Ohio State Law J.*, vol. 78, no. 5, pp. 1217–1241, 2017.
5. N. Bostrom, *Superintelligence: Paths, Dangers, Strategies*. Oxford, U.K.: Oxford Univ Press., 2014.
6. M. R. Calo, "Open robotics," *Maryland Law Rev.*, vol. 70, no. 3, pp. 571–613, 2011.
7. S. Chesterman, *We, the Robots? Regulating Artificial Intelligence and the Limits of the Law*. Cambridge, U.K.: Cambridge Univ. Press, 2021.
8. D. Collingridge, *The Social Control of Technology*. London, U.K.: Frances Pinter, 1980.
9. "Largest companies by market cap." Companies Market Cap. Accessed: May 2022. [Online]. Available: <https://companiesmarketcap.com>
10. Editorial, "Cambridge Analytica controversy must spur researchers to update data ethics," *Nature*, vol. 555, pp. 559–560, Mar. 2018.
11. U. Gasser and V. A. F. Almeida, "A layered model for AI governance," *IEEE Internet Comput.*, vol. 21, no. 6, pp. 58–62, Nov./Dec. 2017, doi: [10.1109/MIC.2017.4180835](https://doi.org/10.1109/MIC.2017.4180835).
12. M. Hu, "Cambridge Analytica's black box," *Big Data Soc.*, vol. 7, no. 2, 2020, Art. no. 2053951720938091, doi: [10.1177/2053951720938091](https://doi.org/10.1177/2053951720938091).
13. A. Jobin, M. Ienca, and E. Vayena, "The global landscape of AI ethics guidelines," *Nature Mach. Intell.*, vol. 1, no. 9, pp. 389–399, 2019, doi: [10.1038/s42256-019-0088-2](https://doi.org/10.1038/s42256-019-0088-2).
14. L. Munn, "The uselessness of AI ethics," *AI Ethics*, vol. 3, no. 3, pp. 869–877, 2023, doi: [10.1007/s43681-022-00209-w](https://doi.org/10.1007/s43681-022-00209-w).
15. G. Papyshv and M. Yarime, "The state's role in governing artificial intelligence: Development, control, and promotion through national strategies," *Policy Des. Pract.*, vol. 6, no. 1, pp. 79–102, doi: [10.1080/25741292.2022.2162252](https://doi.org/10.1080/25741292.2022.2162252).
16. F. Pasquale, *New Laws of Robotics*. Cambridge, MA, USA: Harvard Univ. Press, 2020.
17. J. Schneider, R. Abraham, C. Meske, and J. Vom Brocke, "Artificial intelligence governance for businesses," *Inf. Syst. Manage.*, vol. 40, no. 3, pp. 229–249, 2023, doi: [10.1080/10580530.2022.2085825](https://doi.org/10.1080/10580530.2022.2085825).
18. M. Srikumar, R. Finlay, G. Abuhamad et al., "Advancing ethics review practices in AI research," *Nature Mach. Intell.*, vol. 4, no. 12, pp. 1061–1064, 2022, doi: [10.1038/s42256-022-00585-2](https://doi.org/10.1038/s42256-022-00585-2).
19. A. Theodorou and V. Dignum, "Towards ethical and socio-legal governance in AI," *Nature Mach. Intell.*, vol. 2, no. 1, pp. 10–12, 2020, doi: [10.1038/s42256-019-0136-y](https://doi.org/10.1038/s42256-019-0136-y).
20. "Regional groups of member states," United Nations, San Francisco, CA, USA. 2024. [Online]. Available: <https://www.un.org/dgacm/en/content/regional-groups>