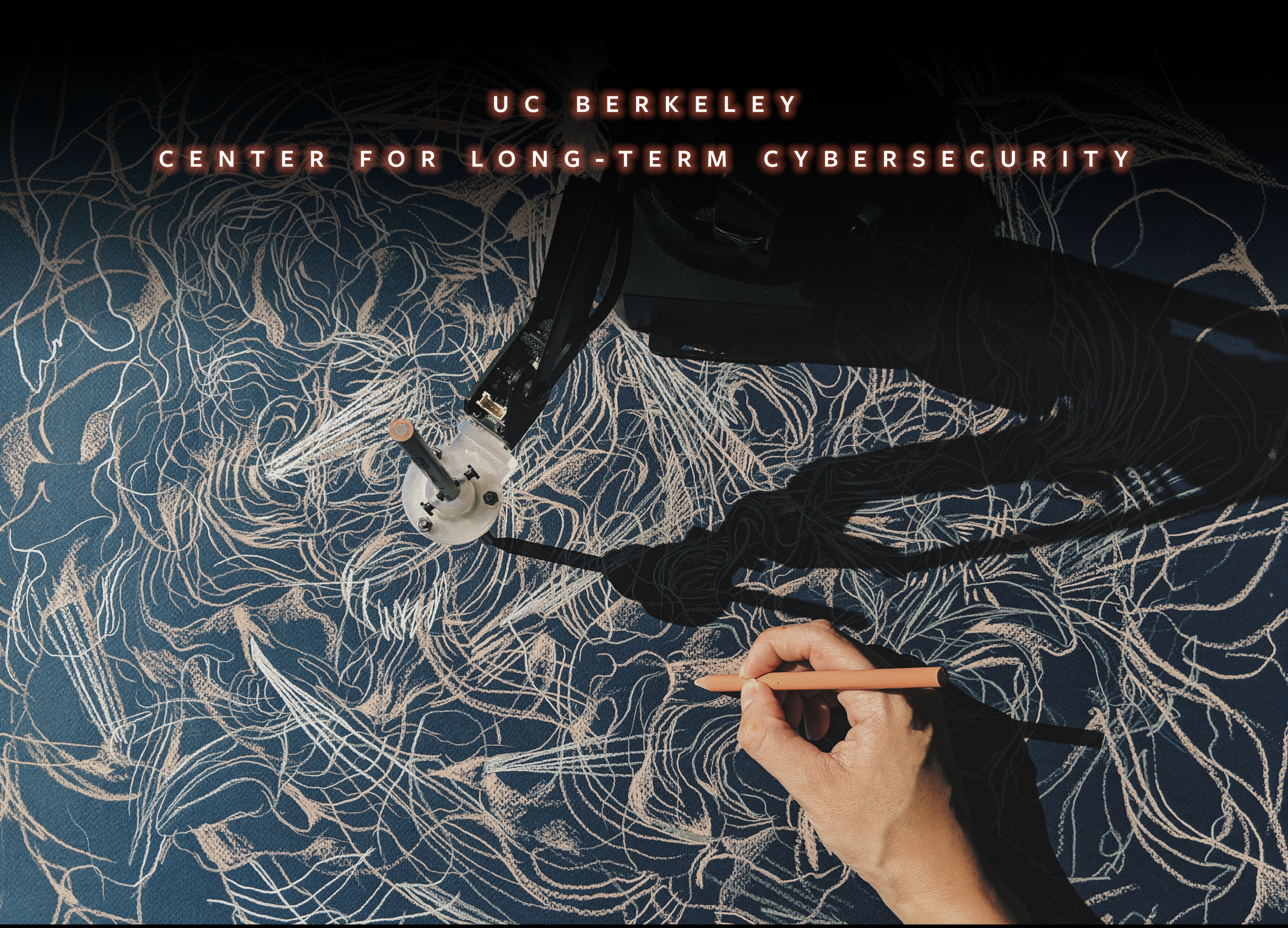
UC BERKELEY
CENTER FOR LONG-TERM CYBERSECURITY



CLTC WHITE PAPER SERIES

Decision Points in AI Governance

THREE CASE STUDIES EXPLORE EFFORTS TO OPERATIONALIZE AI PRINCIPLES

JESSICA CUSSINS NEWMAN

CLTC WHITE PAPER SERIES

DECISION POINTS IN AI GOVERNANCE

THREE CASE STUDIES EXPLORE EFFORTS TO OPERATIONALIZE AI PRINCIPLES

JESSICA CUSSINS NEWMAN

CLTC
Center for Long-Term Cybersecurity
UC Berkeley

UC BERKELEY

CENTER FOR LONG-TERM CYBERSECURITY

Contents

Acknowledgments

The Center for Long-Term Cybersecurity (CLTC) would like to thank the following individuals for contributing their ideas and expertise to this report through interviews, conversations, feedback, and review: Jared Brown, Miles Brundage, Peter Cihon, Allan Dafoe, Cyrus Hodes, Eric Horvitz, Will Hunt, Daniel Kluttz, Yolanda Lannquist, Herb Lin, Nick Merrill, Nicolas Miailhe, Adam Murray, Brandie Nonnecke, Karine Perset, Toby Shevlane, and Irene Solaiman.

Special thanks also to Steven Weber, Ann Cleaveland, and Chuck Kapelke of CLTC for their ongoing support, feedback, and contributions.

CLTC would also like to thank the Hewlett Foundation for making this work possible.

About the cover Image: This image depicts the creation of "Artifact 1," by Sougwen Chung, a New York-based artist and researcher. Artefact 1 (2019) explores artistic co-creation and is the outcome of an improvisational drawing collaboration with a custom robotic unit linked to a recurrent neural net trained on Ms. Chung's drawings. The contrasting colors of lines denote those marks made by the machine and by Ms. Chung's own hand. Learn more at https://sougwen.com/.

Executive Summary

Since 2016, dozens of groups from industry, government, and civil society have published "artificial intelligence (AI) principles," frameworks designed to establish goals for safety, accountability, and other goals in support of the responsible advancement of AI. Yet while many AI stakeholders have a strong sense of ""what" is needed, less attention has focused on "how" institutions can translate strong AI principles into practice.

This paper provides an overview of efforts already under way to resolve the translational gap between principles and practice, ranging from tools and frameworks to standards and initiatives that can be applied at different stages of the AI development pipeline. The paper presents a typology and catalog of 35 recent efforts to implement AI principles, and explores three case studies in depth. Selected for their scope, scale, and novelty, these case studies can serve as a guide for other AI stakeholders — whether companies, communities, or national governments — facing decisions about how to operationalize AI principles. These decisions are critical because the actions AI stakeholders take now will determine whether AI is safely and responsibly developed and deployed around the world.

Microsoft's AI, Ethics and Effects in Engineering and Research (AETHER) Committee: This case study explores the development and function of Microsoft's AETHER Committee, which has helped inform the company's leaders on key decisions about facial recognition and other AI applications. Established to help align AI efforts with the company's core values and principles, the AETHER Committee convenes employees from across the company into seven working groups tasked with addressing emerging questions related to the development and use of AI by Microsoft and its customers.

The case study provides lessons about:

- How a major technology company is integrating its AI principles into company practices and policies, while providing a home to tackle questions related to bias and fairness, reliability and safety, and potential threats to human rights and other harms.
- Key drivers of success in developing an AI ethics committee, including buy-in and participation from executives and employees, integration into a broader company culture of responsible AI, and the creation of interdisciplinary working groups.

OpenAI's Staged Release of GPT-2: Over the course of nine months in 2019, OpenAI, a San Francisco-based AI research laboratory, released a powerful AI language model in stages —

rather than all at once, the industry norm — in part to identify and address potential societal and policy implications. The company's researchers chose this "staged release" model as they were concerned that GPT-2 — an AI model capable of generating long-form text from any prompt — could be used maliciously to generate misleading news articles, impersonate others online, automate the production of abusive content online, or automate phishing content.

The case study provides lessons about:

- Debates around responsible publication norms for advanced AI technologies.
- How institutions can use threat modeling and documentation schemes to promote transparency about potential risks associated with their AI systems.
- How AI research teams can establish and maintain open communication with users to identify and mitigate harms.

The OECD AI Policy Observatory: In May 2019, 42 countries adopted the Organisation for Economic Co-operation and Development (OECD) AI Principles, a legal recommendation that includes five principles and five recommendations related to the use of AI. To ensure the successful implementation of the Principles, the OECD launched the AI Policy Observatory in February 2020. The Observatory publishes practical guidance about how to implement the AI Principles, and supports a live database of AI policies and initiatives globally. It also compiles metrics and measurement of global AI development and uses its convening power to bring together the private sector, governments, academia, and civil society.

The case study provides lessons about:

- How an intergovernmental initiative can facilitate international coordination in implementing AI principles, providing a potential counterpoint to "AI nationalism."
- The importance of having several governments willing to champion the initiative over numerous years; convening multistakeholder expert groups to shape and drive the agenda; and investing in significant outreach efforts to global partners and allies.

The question of how to operationalize AI principles marks a critical juncture for AI stakeholders across sectors. Getting this right at an early stage is important because technological, organizational, and regulatory lock-in effects are likely to make initial efforts especially influential. The case studies detailed in this report provide analysis of recent, consequential initiatives intended to translate AI principles into practice. Each case provides a meaningful example with lessons for other stakeholders hoping to develop and deploy trustworthy AI technologies.

Introduction

Research and development in artificial intelligence (AI) have led to significant advances in natural language processing, image classification and generation, machine translation, and other domains. Interest in the AI field has increased substantially, with 300% growth in the volume of peer-reviewed AI papers published worldwide between 1998 and 2018, and over 48% average annual growth in global investment for AI startups.[1] These advances have led to remarkable scientific achievements and applications, including greater accuracy in cancer screening and more effective disaster relief efforts. At the same time, growing awareness of the significant safety, ethical, and societal challenges stemming from the advancement of AI has generated enthusiasm and urgency for establishing new frameworks for responsible governance.

The emerging "field" of AI governance — interconnected with such fields as privacy and data governance — has moved through several stages over the past four years. The first stage, which began most notably in 2016, has been characterized by the emergence of AI principles and strategies enumerated in documents published by governments, firms, and civil-society organizations to clarify specific intentions, desires, and values for the safe and beneficial development of AI. Much of the AI governance landscape thus far has taken the form of these principles and strategy documents, at least 84 of which were in existence as of September 2019.[2]

The second stage, which initially gained traction in 2018, was characterized by the emergence of efforts to map this proliferation of AI principles[3] and national strategies[4] to identify divergences and commonalities, and to highlight opportunities for international and multistakeholder collaboration. These efforts have revealed growing consensus around a number of central themes, including privacy, accountability, safety and security, transparency and explainability, fairness and non-discrimination, human control of technology, professional responsibility, and promotion of human values.[5]

The third stage, which largely began in 2019, has been characterized by the development of tools and initiatives to transform AI principles into practice. While the first two stages helped shape an international AI "normative core," there has been less consensus about how to achieve the goals defined in the principles. Much of the debate about AI governance has focused on 'what' is needed, as laid out in the principles and guidelines, but there has been less focus on the 'how,' the practices and policies needed to implement established goals. This paper argues that the question of how to operationalize AI principles and strategies is one of the key decision points that AI stakeholders face today, and offers examples that may help AI stakeholders navigate the challenging decisions they will face.

Efforts are already under way to resolve the translational gap between principles and practice, ranging from tools and frameworks to standards and initiatives that can be applied at different stages of the AI development pipeline.[6] A partial and non-exhaustive list of such efforts can be found in the tables below, organized into six categories: Technical Tools; Oversight Boards and Committees; Frameworks and Best Practices; Standards and Certifications; Regulations; and Institutions and Initiatives. The typology provides examples of the kinds of efforts now under way, specifically for AI research and applications. Efforts included here are well represented in the literature, but were not independently vetted for efficacy or adoption rates.

EXISTING EFFORTS TO TRANSLATE PRINCIPLES TO PRACTICE

Technical Tools

Name	Description
The Equity Evaluation Corpus[7]	A database consisting of thousands of sentences chosen to help determine when automatic systems demonstrate biases toward certain races and genders
InterpretML[8]	An open-source library developed by Microsoft for training interpretable machine-learning models and explaining black-box systems
The Adversarial Robustness 360 Toolbox[9]	A library supporting developers and researchers in defending machine-learning models against adversarial threats
The AI Fairness 360 Toolkit[10]	A toolkit of metrics to check for unwanted bias in datasets and machine-learning models, as well as algorithms to mitigate such bias
The TensorFlow Privacy library[11]	A library to help train machine-learning models with differential privacy
The Accenture AI Fairness Toolkit[12]	A tool to help companies detect and eliminate gender, racial, and ethnic bias in AI systems

Oversight Boards and Committees

Name	Description
Microsoft AETHER Committee[13]	A committee composed of seven working groups within Microsoft to focus on proactive formulation of internal policies for responding to specific AI-related issues in a responsible way
Google Responsible Innovation Team[14]	An interdisciplinary group that handles day-to-day operations and initial assessments about the ethics of Google AI products
DeepMind Fellows[15]	A group of independent advisors who help provide critical feedback and guidance to the AI company DeepMind
Axon AI and Policing Technology Ethics Board[16]	An independent external advisory board that provides expert guidance on the development of AI products and services

Frameworks and Best Practices

Name	Description
Model Cards for Model Reporting[17]	Short documents accompanying trained machine-learning models that provide benchmarked evaluation of performance in a variety of relevant conditions
Datasheets for Datasets[18]	A process for documenting the motivation, composition, and collection process for datasets, and their recommended uses
Staged Release[19]	A publication methodology intended to minimize misuse potential (most notably used by OpenAI to release progressively larger versions of the language model GPT-2)
Scoping, Mapping, Artifact Collection, Testing, and Reflection (SMACTR)[20]	A framework for algorithmic auditing that supports AI system development from end to end, to be applied throughout the development life cycle
The Assessment List of the Ethics Guidelines for Trustworthy AI[21]	A resource developed by the European Commission's High Level Expert Group on Artificial Intelligence to operationalize key requirements for ethical AI and offer guidance on practical implementation

The a3i Trust-in-AI Framework[22]	A model to help organizations integrate safety, security, and explainability into the design of AI systems
The AI-RFX Procurement Framework[23]	A set of templates to support industry practitioners in the procurement of AI systems
Algorithmic Accountability Policy Toolkit[24]	A toolkit from the AI Now Institute that includes resources for advocates working to uncover where algorithms are being used and to improve transparency and accountability mechanisms
The Oversight Toolkit for Boards of Directors[25]	A reference developed by the World Economic Forum to help boards of directors advance the beneficial use of AI in their companies
Algorithmic Impact Assessments[26]	A framework developed by the AI Now Institute to help public agencies assess automated decision systems and ensure public accountability
Human Rights Impact Assessments[27]	A process for assessing and addressing the impacts of products or activities – including AI technologies – on human rights
PST Framework[28]	Short for "predictability, computability, and stability," a framework developed by UC Berkeley Professor Bin Yu that promotes responsibility throughout the data science and machine-learning life cycle

Standards and Certifications

Standards Body	Standard / Certification Name
The International Organization for Standardization (ISO) (ISO/IEC JTC 1)[29]	ISO/IEC 20546:2019: Information technology — Big data — Overview and vocabulary, which provides a terminological foundation for big data-related standards
	ISO/IEC TR 20547-2:2018: Information technology — Big data reference architecture — Part 2: Use cases and derived requirements, which provides examples of big data use cases with application domains and technical considerations

	ISO/IEC TR 20547-5:2018: Information technology — Big data reference architecture — Part 5: Standards roadmap, which describes big data relevant standards in existence and under development, along with priorities for future big data standards development based on gap analysis
The Institute of Electrical and Electronics Engineers (IEEE)[30]	Ethics Certification Program for Autonomous and Intelligent Systems (ECPAIS), which develops metrics and processes toward the implementation of a certification methodology addressing transparency, accountability, and algorithmic bias
	The IEEE P7000™ Standards Series, which includes 14 AI-related standards

Regulations

Jurisdiction	Description
European Union, China, California	Data rights and privacy laws, including the EU General Data Protection Regulation (GDPR), China's Personal Information Security Specification, and the California Consumer Privacy Act (CCPA)
Cities, including San Francisco and Oakland, California and Somerville, Massachusetts; and States, including California, Oregon, and New Hampshire	Restrictions and bans on government and law enforcement's use of facial recognition technologies
California	Restrictions and legal liability for the use of 'deepfakes' and other synthetic media of political candidates or for nonconsensual sexual content
California	A bot disclosure law, making it illegal to use a bot with the intention of misleading another person for commercial or political purposes
New York, Vermont	State Commissions to study AI and propose regulations

Institutions and Initiatives

Name	Description
OECD AI Policy Observatory	Builds on the momentum of the OECD's Recommendation on Artificial Intelligence to facilitate dialogue and provide multidisciplinary, evidence-based policy analysis
Partnership on AI's ABOUT ML	An initiative to enable organizations to operationalize responsible AI by increasing transparency and accountability with machine-learning system documentation
The Global Partnership for AI (GPAI)	An international and multistakeholder forum to monitor and debate the policy implications of AI globally, initiated by French president Emmanuel Macron and Canadian prime minister Justin Trudeau
IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems	An initiative that aims to move "from principles to practice" with standards projects, certification programs, and global consensus-building to inspire the ethically aligned design of autonomous and intelligent technologies
Global Governance of AI Forum (GGAF), World Government Summit	The GGAF gathers multilateral institutions and multi-stakeholder experts to discuss the global governance of AI and offers a unique yearly coordination mechanism
AI for Good Global Summit	An annual summit focused on accelerating the United Nations Sustainable Development Goals, described as the "leading action-oriented, global & inclusive United Nations platform on AI"

In addition to the resources named in the tables, countless published papers have helped define best practices for facilitating privacy, security, safety, fairness, and explainability throughout AI development.

The emergence of these resources and initiatives resulted from growing acknowledgment that researchers, companies, and policymakers can, and should, do more to mitigate known and unknown risks related to artificial intelligence. We have reached a crucial juncture in the decades-long history of technological development. AI technologies are now being deployed in critical functions and for consequential ends, including the generation of synthetic media (AI-generated images, text, audio, and video), automated decision-making in the military and other high-stakes environments, and the advancement of powerful tools for security and surveillance. However, it is increasingly understood that AI systems make significant errors. They can be tricked and misused, can make damaging mistakes, and may result in unintended consequences at massive scales. There is also uncertainty about how to establish practices to make reasonable tradeoffs between occasionally conflicting goals.

As a result of past mistakes — including the abuse of consumer data and the pursuit of controversial military and foreign contracts — the U.S. technology industry has come under pressure to repair lost trust with consumers and employees. Less clear is whether this pressure, combined with the public expression of principles from influential AI stakeholders, can mature into the implementation of new processes, standards, and policies. The actions AI stakeholders take now to achieve these goals will help determine whether the full upside potential of these technologies can be realized.

AI DECISION POINTS

Many decisions are made throughout the life cycle of an AI system, from which training data to use (if any), to whether and how to test for the robustness of a model against attack. All of these decisions affect the reliability, security, and impacts of the system. Transparency about how and why these decisions were made has become an important element of accountability. AI decision-makers have to consider tradeoffs and priorities, and weigh which decisions are likely to have an outsized impact.

The concept of "AI decision points" introduced in this paper refers to concrete actions taken by an AI stakeholder (organization, company, government, employees, etc.) that were not predetermined by existing law or practice, and that mark a meaningful shift in behavior from previous practice, for the purpose of shaping the development and use of AI. These decisions are catalysts of broader inflection points leading to new strategies and opportunities for the organization.

Because AI decision points reflect efforts to actively shape the field, they are poised to have a disproportionately large influence on the future trajectory of AI. Tracking decision points can provide insight into existing tensions and challenges, and into how the field is evolving to address them. Identifying decision points can also be a useful way to focus governance efforts, as it narrows a crowded solution space and highlights opportunities to make influential decisions.

AI technologies are valuable in part because of their ability to automate elements of complex decision-making. This utility can at times obscure the importance of human decisions in shaping the design and use of AI technologies. However, AI systems are shaped by countless decisions, related not only to technical design, but also to human and institutional preferences. Attention to decision points underscores that specific AI trajectories are not inevitable, and there are opportunities to make decisions that will support safer and more responsible development and use.

This analysis is primarily relevant to people interested in shaping future trajectories of AI, including AI researchers, policymakers, and industry executives. That is not to suggest that there is a single right answer, or that decisions should be made in isolation or only by people in positions of power. Indeed, the decisions described in the case studies were all made through consultative and iterative processes. AI technologies impact many people, but can have disparate effects across different communities. Legitimate governance efforts include engagement with diverse stakeholders as well as impacted communities.

The case studies in this paper are intended to highlight key levers that are currently being used to shape the future of AI, to provide examples from which the field can draw, and to support further analysis of the effectiveness and desirability of such efforts as tools of AI governance. There is no need for AI stakeholders to start from scratch in their endeavor to operationalize AI principles. No single actor can accomplish the mitigation of AI threats in isolation; stakeholders need to coordinate and cooperate with each other, which is much easier with improved collective understanding of ongoing efforts.

On the Need to Operationalize AI Principles

Decisions about how to operationalize AI principles and strategies have potential to shift the AI landscape toward a more safe and responsible future trajectory. A literature review of recent developments in AI and AI governance reveals growing awareness of the need to support principled implementation efforts.[31] It has become difficult for AI stakeholders to ignore the many AI principles and strategies in existence. Even companies and organizations that have not defined their own principles may be expected or required to adhere to those adopted by governments. This growing "universality" will lead to increased pressure to establish methods to ensure AI principles and strategies are realized. Meaningful implementation efforts are likely to be critical in maintaining trust with users and the public. Due to technological, organizational, and regulatory lock-in effects, early efforts — those that fill governance gaps to establish new standards and best practices — are likely to be especially influential.

The goals defined within the various AI principles and strategies are both practical and ambitious, and often build upon other human rights frameworks, rights and laws, and non-rights perspectives. These documents have emerged from all sectors, including governments, corporations, and civil society. The stated purpose of many of the principles is to forge trust: between governments and citizens, between corporations and consumers or users, and between AI researchers and the general public. However, in the wake of the "techlash" — a term used to describe growing public animosity toward large technology companies — earning trust is not straightforward. Trust is rather reserved for those entities that provide compelling values and motivations for their work and back them up with meaningful actions.

Pledges of intention, made explicit through principles and strategies, can be extremely valuable. They offer clarity and guidance about paths to pursue, and provide insight into the priorities of firms and governments. They can also help to hold actors accountable. For example, Google's AI Principles include a list of AI applications the company will *not* pursue, including weapons, surveillance technologies that violate international norms, or any technologies that contravene international law and human rights.[32] Nonetheless, people are growing unsatisfied with promises alone, and are calling out companies and governments for failing to act on their principles.[33 / 34]

While there is a logical trajectory for stakeholders to move from "principles to practice" (while facilitating an ongoing feedback loop between the two), this transition is only possible with sufficient agreement about what safe and responsible AI development and use entails. AI principles do not have to be universally accepted or non-controversial; cultural and political dynamics necessarily include variability. Rather, how principles, strategies, and guidelines are translated into action will need to accord with international laws and standards, and with unique national, local, and organizational needs. Implementation efforts are therefore most easily understood with sufficient context about their origin and scope.

The case studies below attempt to provide such specificity. Each case can be understood as a meaningful example (though not necessarily an ideal model that all others should replicate), offering relevant lessons for AI stakeholders hoping to shape the future of AI. They showcase how technology companies and governments are grappling with key decisions in AI governance, and highlight some of the successes and challenges faced by key AI stakeholders around the world.

Case Studies

The following three case studies provide analysis of recent, consequential decisions that were made to operationalize AI principles in order to support safe and responsible AI development. The case studies highlight the example of an ethics and impact advisory committee re-shaping engineering and research practices at a top technology company; an AI research laboratory that experimented with staged release and other accountability measures for the dissemination of a potentially harmful AI model; and the first prominent example of intergovernmental and multistakeholder AI coordination through a common global framework.

These case studies were selected based upon their scope, scale, and novelty. Many of the existing translational efforts described in the tables on pages 4–8 are narrowly focused on a particular AI challenge, such as reducing the threat from adversarial attacks or mitigating algorithmic bias. These tools are critical, but this paper is concerned with efforts of a broader scope. Each example described below simultaneously addresses numerous interconnected issues that affect the likelihood of ensuring safe and responsible AI development. The scale of each of these examples is also notable. While many of the tools listed previously are open-source and could be used broadly, it is not clear whether many of them have been widely adopted. However, the three examples below represent shifts in practices and polices that were made across entire companies and organizations, with evidence of spillover effects to other parts of the AI ecosystem already present.

Lastly, these case studies were selected because they represent novel shifts in well-established behaviors. Microsoft's decision to adopt an ethics and effects advisory committee marked a reportedly first-of-its-kind effort to formalize review of societal impact throughout the lifecycle of AI technologies at the core of a company's business model. OpenAI's decision to adopt accountability measures and the staged release of an AI model represented a marked shift from the open publication norms in AI and machine-learning communities. And the OECD's decision to establish an intergovernmental hub for AI governance serves as a counterpoint to AI nationalism and rhetoric about an "AI race" between nations. The novelty of these examples makes them more impactful because they signify inflection points, or significant changes, in the AI landscape. Moreover, each case study explores the implementation of AI principles for different kinds of key AI stakeholders: a large technology company, an AI research lab, and governments around the world.

CASE STUDY I

CAN AN AI ETHICS ADVISORY COMMITTEE HELP CORPORATIONS ADVANCE RESPONSIBLE AI?

Understanding the Role of the Microsoft AETHER Committee

INSIGHTS

- Large multinational companies have an outsized impact on trends in AI development and deployment, but have not universally adopted new practices or oversight committees to help ensure their technologies will be beneficial.
- Microsoft aims to be a leader in responsible AI, and has established the AETHER Committee with the intention of operationalizing the company's AI principles into its engineering and business practices.
- The AETHER Committee is facilitating internal deliberation about controversial use-cases, providing channels for concerned employees, and incentivizing research in the areas of its working groups, including safety, security, and accountability.
- The AETHER Committee attributes its success in part to executive-level support, regular opportunities for employee and expert engagement, and integration with the company's legal team.

How the AETHER Committee Emerged and How it Works

In a March 2018 email to all employees, Satya Nadella, the chief executive officer of Microsoft, described the importance of AI innovation to the long-term success of the company, noting that ensuring responsibility was critical as the technology progressed.[35] He announced that Brad Smith, president of Microsoft, and Harry Shum, executive vice president of the company's AI and Research group, would be establishing the AI and Ethics in Engineering and Research (AETHER) Committee. He described the Committee as a way to bring together senior leaders from throughout the company to build proactive internal policies and address specific issues raised. He told employees, "AETHER will ensure our AI platform and experience efforts are deeply grounded within Microsoft's core values and principles and benefit the broader society." By early 2019, the meaning of the acronym "AETHER" was expanded to stand for AI, Ethics and Effects in Engineering and Research.[36]

At a time when numerous technology firms face pressure from the growing "techlash," several companies have established AI ethics committees of various kinds to try to manage growing challenges. However, these committees are still viewed with some suspicion, and in some cases have been called out as "AI ethics-washing."[37] Do such committees reflect calculated public relations moves, or can they be effective for translating AI principles into practice? The Microsoft AETHER Committee represents a novel experiment to restructure practices across the company's AI engineering and research teams, and may provide lessons on challenges and opportunities for other organizations. This analysis is based upon interviews and conversations with Microsoft executives and employees, a review of public documentation and media, and a 2019 presentation about AETHER given at UC Berkeley.

In 2018, Microsoft outlined six principles to guide the company's AI development in a book titled *The Future Computed: Artificial Intelligence and its role in society*.[38] The principles laid out in the book include fairness, reliability, privacy and security, inclusiveness, transparency, and accountability. On their own, the principles are not particularly noteworthy; they reinforce common AI principles defined by countless organizations. However, while other efforts to institutionalize guiding values for AI into industry have struggled,[39] Microsoft has made an extensive effort to restructure engineering practices and policies around its values, and has had some notable successes. This case study examines how the AETHER Committee (hereinafter referred to as AETHER) is designed, how it functions, and what its impacts have been.

AETHER helps organize internal talent at Microsoft, calling upon employees from different backgrounds to address controversial or complex cases and to proactively develop internal policies for safe and ethical AI development. AETHER members write reports about specific questions, outlining the issues at stake and the costs and benefits of a particular action, and then present the information, along with recommendations, to Microsoft's senior leadership team. These recommendations provide guidance and contribute to practices, policies, and positions at the company. AETHER is primarily an internal advisory committee, but has occasionally consulted with outside experts. AETHER also works alongside another group at Microsoft, the Office of Responsible AI, which is based in the legal department and assists with compliance efforts.

In a keynote address in November 2019, the chair of AETHER, Eric Horvitz, explained that the role of AETHER is to "advise the senior leadership team on policies around sensitive issues when it comes to AI products and services."[40] He added, "it's already had quite a significant effect on gating and guiding Microsoft technologies and how it works with customers in different parts of the world when it comes to these technologies." To date, little information about AETHER or its impact has been publicly disclosed. This case study aims to shed some light on the structure and achievements of the AETHER Committee at a relatively early stage.

Working Groups

AETHER is organized into seven working groups, generally composed of 5–7 people at their core and twenty people in total, including co-chairs (typically top experts in the field,) a core subgroup of committee members, and an expanded subgroup with representatives from every major division in the company. The working groups are dedicated to the following focus areas:

- *Sensitive Uses*, to assess automated decision-making that can have a major impact on people's lives, such as the denial of consequential services, risks to human rights, and risks of harm;
- *Bias and Fairness*, to investigate potential impacts of AI systems on minority and vulnerable populations;
- *Reliability and Safety*, to ensure AI systems function as intended and are robust against adversarial attacks;
- *Human Attention and Cognition*, to monitor algorithmic attention-hacking and abilities of persuasion;
- *Intelligibility and Explanation*, to promote transparency into how machine-learning and deep-learning models process data and reach decisions;
- *Human-AI Interaction and Collaboration*, to study how people can better and more productively engage with AI systems; and
- *Engineering Best Practices*, to recommend education, training, best practices, processes, and tooling to support each stage of the AI system development cycle and ensure that Microsoft teams are equipped and motivated to apply Microsoft principles for responsible AI.

In addition to serving an advisory role, the working groups also address technical challenges and publish externally. For example, the Bias and Fairness working group has looked at Microsoft's facial recognition service to tackle problems associated with gender recognition for women with darker skin. The group developed new tools to probe the system and better understand and address its limitations. Similarly, the working group on Engineering Practices published a paper on AI threat-modeling,[41] and the Intelligibility and Explanation working group developed an open-source Python package, InterpretML, intended to train interpretable machine learning models and help explain black-box systems.[42]

The Sensitive Uses working group specifically has a mission to undertake "analysis and deliberation about AI systems involved in sensitive uses, including automated recommendations that can have deep impact on people's lives." As determining what constitutes a "sensitive use" of AI is not always a straightforward task, Microsoft leans on influential precedents, including the Universal Declaration of Human Rights, the Guiding Principles on Business and Human Rights, and Microsoft's own Human Rights Policy. Concrete examples of sensitive uses include the denial of credit, employment, education, or healthcare services; the use of surveillance systems and other AI systems that pose risks to personal freedoms, privacy, and human rights; and the risk of AI systems creating significant physical or emotional harm.

This working group relays recommendations to Microsoft's senior leadership team, who may agree or disagree with the findings. If a new recommendation is agreed to, the working group can then establish a new policy for the company, providing an important pathway for AETHER to establish precedent. This ability to generate new company policies for Microsoft is one of AETHER's critical functions, as it can lead to the restructuring of engineering or other institutional processes. Greater external transparency about this process will help clarify how many of the recommendations do in fact generate new company policies.

AETHER is also called upon to resolve questions about the implications of Microsoft's AI products and services, and about how to manage sensitive customer requests or uses. For example, AETHER reviews how customers might use (and misuse) Microsoft's AI products. According to Horvitz, this resulted in the company changing its course as early as April 2018.[43] "Significant sales have been cut off. And in other sales, various specific limitations were written down in terms of usage, including 'may not use data-driven pattern recognition for use in face recognition or predictions of this type.'" Horvitz added, "It's been an intensive effort ... and I'm happy to say that this committee has teeth." Evidence to verify his claim can be found in the examples of rejected requests described below.

Impacts and Challenges

Facial recognition technology has been a keen area of focus for AETHER, and contributed to Microsoft's calls for regulation of this emerging use of AI. In April 2019, it was revealed that Microsoft rejected a request from a sheriff's department in California to install facial recognition technology in officers' cars and body cameras because the company determined that to do so would constitute a human rights concern, given the high likelihood of bias against minorities.[44] Microsoft has also rejected requests from foreign governments to install facial recognition on surveillance cameras due to concerns this could suppress freedom of assembly.

Microsoft has, however, allowed many uses of its AI technologies based upon recommendations from AETHER. For example, the company facilitated the use of facial recognition technology within an American prison after AETHER assessed that its uses would be limited and likely to improve safety conditions. More controversially, Microsoft has supported a facial recognition company in Israel called AnyVision, which has been criticized for identifying and tracking Palestinians in the West Bank.[45] In November 2019, following concerns about potential human rights abuses, Microsoft hired former United States Attorney General Eric Holder to investigate whether AnyVision appropriately complies with the company's principles for the use of AI and facial recognition technology.[46]

Other efforts will also test Microsoft's commitment to principled implementation of AI in the coming years. In October 2019, in a surprise upset to frontrunner Amazon, Microsoft won an historic $10 billion contract with the Department of Defense (DoD) to transform the military's cloud computing systems. Google dropped out of the competition for the contract in 2018 in part because the work was determined to go against the company's AI principles, which state they will not contribute to the use of AI in weaponry. Notably, Microsoft's AI principles do not preclude such uses, as long as the systems are reliable, safe, and accountable.

Microsoft has strong convictions about the importance of supporting the U.S. military with its technologies, but has not shied away from addressing the sensitivities associated with its decisions.[47] AETHER's Sensitive Uses working group reportedly defined a policy on the company's sale of AI technologies to the Department of Defense following an executive retreat on the matter.

AETHER Chair Eric Horvitz has further defined key challenges related to the use of AI in military applications, providing greater insight into how the company is thinking about its role and responsibilities.[48] Horvitz said in a November 2019 keynote address for the Bulletin of the Atomic Scientists that "inescapable errors from AI systems" must be taken into consideration and that efforts should be made to allow time for "human reflection, input, and intervention." Horvitz did not say that humans should always remain in the loop, but he did say that removing people from these positions of oversight should only be done with "wisdom and caution."

He also called for consideration of an array of challenges related to AI, including the impossibilities of fully testing AI capabilities in realistic, operational settings; the rise of unexpected behaviors in the complexities of interactions among AI systems; preparation for adversarial attacks on AI systems; vigilance against new forms of persuasion and deception; collaboration with potential adversaries to minimize instabilities and facilitate human oversight; investment in human-AI interaction technologies; and vigilance for the assertion of ethical principles for AI that changes the nature of war. These challenges do not directly translate to operational practices however, and Microsoft will likely face a higher degree of scrutiny over its uses of AI in the coming years.

Microsoft has been providing technologies to the Department of Defense for more than 40 years. However, not all Microsoft employees have been on board with the company's willingness to support the U.S. military. A group called Microsoft Workers 4 Good, whose mission is "to empower every worker to hold Microsoft accountable to their stated values," has called on Microsoft leadership to end certain contracts. For example, in February 2019, the group sent a letter calling on Brad Smith and Satya Nadella to end a contract through which the company would provide its HoloLens augmented reality technology to the U.S. Army to support war fighting.

The letter stated, "We are alarmed that Microsoft is working to provide weapons technology to the U.S. Military, helping one country's government 'increase lethality' using tools we built. We did not sign up to develop weapons, and we demand a say in how our work is used."[49] Nadella defended the company's decision, stating, "We made a principled decision that we're not going to withhold technology from institutions that we have elected in democracies to protect the freedoms we enjoy." Smith has also defended the contract with the Defense Department, arguing, "All of us who live in this country depend on its strong defense. . . . We want the people of this country and especially the people who serve this country to know that we at Microsoft have their backs. They will have access to the best technology that we create." AETHER's Sensitive Uses working group will continue to assess potentially controversial use cases that emerge in this arena.

Features of Success

AETHER has managed to produce new research insights, establish new company policies, and help inform decisions about sensitive uses of Microsoft's AI technology with ongoing support from executives and employees within the company. It has succeeded in part because it is only one piece of a broader ecosystem intended to facilitate a culture of responsible AI development. Other Microsoft efforts that reinforce this commitment include a group called Fairness, Accountability, Transparency and Ethics in AI (FATE), which consists of nine researchers "working on collaborative research projects that address the need for transparency, accountability, and fairness in AI." Additionally, in late 2019, following an AETHER recommendation, Microsoft became the first company in the world to launch a Responsible AI Standard, which is required and informs AI development throughout a system's lifecycle. In 2020, in collaboration with AETHER's Bias and Fairness working group and a group of nearly 50 engineers from numerous technology companies, Microsoft developed an AI ethics checklist for engineers to use throughout the product development process.[50] The company has also added a responsible AI module to educational materials that are required for all employees, and implemented a Responsible AI Champions program, which trains people to be champions for safe and trustworthy AI within their division.

AETHER has several features that make it effective. The inclusion of top executives on the committee signals its perceived importance to the overall mission of the company. AETHER also facilitates input; for example, the committee has established an "Ask AETHER" phone line, which any Microsoft employee can use to raise a concern about an AI technology they are working on or have seen in development or use. Additionally, AETHER's interdisciplinary nature has helped make it more inclusive and far-reaching. Lastly, AETHER helps Microsoft engage proactively with external AI policy developments. For example, Microsoft has repeatedly called for government regulation of facial recognition technology.[51]

The development of a self-regulating body may be seen at least partially as an attempt to prevent external regulation of the company's practices. However, this does not appear to be the primary motivating factor; in addition to the more recent work on the regulation of facial recognition, Microsoft has supported state and federal privacy regulations for years.[52] The company seems to have realized that its internal pivot to expand its focus on AI[53] — and to reorganize the company to integrate AI throughout its divisions and products[54] — means that consumers must have trust in those technologies in order to trust Microsoft. As Microsoft President Brad Smith and Microsoft Senior Director of Communications and External Relations, Carol Ann Browne, wrote in the 2019 book *Tools and Weapons*:

> These issues are bigger than any single person, company, industry, or even technology itself. They involve fundamental values of democratic freedoms and human rights. The tech sector was born and has grown because it has benefited from these freedoms.

> We owe it to the future to help ensure that these values survive and even flourish long after we and our products have passed from the scene.[55]

The decision to establish AETHER stands out because it provides a clear signal to employees, users, clients, and partners that Microsoft intends to hold its technology to a higher standard. AETHER shows one pathway by which companies can empower employees to voice concerns and work toward new company practices and policies supporting the responsible development and use of AI. AETHER is poised to have an outsized effect on the trajectory of AI because it has reshaped a major AI company's practices for vetting and monitoring the AI systems it builds and sells. As of April 2020, this committee-based structure appears to be unique among AI technology companies. Moreover, to date, it has not received significant outside attention because Microsoft has only spoken about it on occasion and has not yet released significant public materials about its processes, though it may do so in the future. As other companies grapple with AI engineering design decisions and the implementation of AI principles, the AETHER Committee provides a valuable example from which other organizations can learn.

CASE STUDY II

CAN SHIFTING PUBLICATION NORMS FOR AI RESEARCH REDUCE AI RISKS?

Accountability Measures and the Staged Release of an Advanced Language Model

INSIGHTS

- Researchers and organizations will increasingly face decisions about how to responsibly publish AI research that could be misused or cause unintentional harm.
- This decision will rarely be a dichotomy between "publish" and "don't publish." "Staged release" is one method along the spectrum as it allows the publication of AI research or technologies over time.
- The test case revealed that there is conflicting evidence about the effectiveness of restraint in keeping a technology from being misused. However, taking the time to partner with other organizations and stakeholders to conduct research into the uses and impacts of a new technology may be valuable.

- Responsible disclosure is not merely about the degree of openness, but also about the implementation of accountability measures, including the use of documentation efforts, discussion of potentially harmful uses and impacts in research papers, and facilitating communication prior to and following the release of new models.

Making Technological (and Cultural) Waves

On February 14, 2019, San Francisco-based AI research laboratory OpenAI announced it had developed an unsupervised language model — a statistical tool that finds patterns in human language and can be used to predict text or audio — capable of generating long-form text from any prompt it received.[56] The model, called GPT-2, came less than a year after the release of GPT, another language model that performed well across a variety of tasks, including tests of reading comprehension, translation, and sentiment analysis. GPT-2 is the next iteration of that model, but was trained on 10X the amount of data (eight million web pages) and has 10X the parameters (1.5 billion). The model was initially trained to predict the next word within a given set of text, but was eventually able to generate many sentences of synthetic text on any topic. The technological advancement was noteworthy in part because the model does not rely upon supervised learning on task-specific datasets, but is capable of learning language processing tasks — machine translation, reading comprehension, and summarization — without explicit supervision.

Beyond the technological achievement, the launch of this powerful AI language model was noteworthy for how OpenAI's research team chose to release it. In what was called "an experiment in responsible disclosure," OpenAI decided not to release the complete trained model, but instead to release a much smaller and simpler model (124 million parameters)[57] along with a technical paper.[58] Over the course of nine months, the organization carried out a "staged release," with progressively larger models released throughout 2019, in May (355 million parameters), August (774 million parameters), and November (1.5 billion parameters, the largest model). The company used the time in between releases to prepare research and documentation exploring the societal and policy implications of the technology alongside the release of the technical papers.

The move sparked debate among AI researchers, who have largely embraced open publishing norms.[59] For example, it has become common for AI researchers to publish early papers on arXiv, a free, open, and non-peer-reviewed publication platform for scientific papers. Funding requirements and job performance metrics have additionally incentivized teams to publish quickly and frequently. The culture around open disclosure in AI can in some regards be viewed as an extension of debates within computer security, where openness is often explicitly utilized for security purposes by allowing the discovery a greater number of software vulnerabilities.[60] Researchers at OpenAI are concerned that this may be a problematic baseline for dual-use research, where the risks of misuse and unintended consequences may outweigh the benefits. Some researchers have suggested that, in certain instances, the AI field should take lessons from domains that exhibit a greater degree of caution in publication than computer security, such as biosecurity and nuclear security.[61]

This case study explores the reasons behind OpenAI's decision to use a staged release for GPT-2, the reaction the company received, and some examples of how norms around responsible disclosure of advanced AI models appear to have shifted. The analysis was based upon interviews, presentations, and feedback from OpenAI employees, as well as a review of public documentation and media.

The Rationale

In April 2018, OpenAI released the OpenAI Charter, which explained that the company's mission is "to ensure that artificial general intelligence (AGI)—by which we mean highly autonomous systems that outperform humans at most economically valuable work—benefits all of humanity." Tenets of the charter include the broad distribution of benefits, long-term safety, technical leadership, and cooperative orientation. The charter also states: "We are committed to providing public goods that help society navigate the path to AGI. Today this includes publishing most of our AI research, but we expect that safety and security concerns will reduce our traditional publishing in the future."[62] In other words, responsible disclosure is a core principle for OpenAI, and the staged release of GPT-2 was an effort to implement this value into organizational practice. The leaders of OpenAI do not take the charter lightly; an in-depth report about the company described the charter as "sacred", informing all performance reviews, strategies, and actions.[63]

The decision to only release a smaller version of GPT-2 at the outset stemmed from concern that the model could be used maliciously, primarily for the generation of scalable, customizable, synthetic media for political or economic gains. For example, OpenAI noted that its model could be used to generate misleading news articles, impersonate others online, automate the production of abusive content online, and automate phishing content. OpenAI wanted to give people more time to adapt and react; they wanted researchers to have more time to work on mitigating risks, and for the public to realize that greater diligence may be required to discern what is true.

The company's concern was not unjustified; AI technology had previously been used for "deepfakes" — where someone's face is inserted into existing video content — for synthetic pornography and content intended to undermine political figures. OpenAI's research and policy team believed caution was warranted in this case because even the first language model they released was able to generate text that seemed relatively authentic. For example, when prompted with two sentences — "A train carriage containing controlled nuclear materials was stolen in Cincinnati today. Its whereabouts are unknown"[64] — the model, on the first try, wrote seven additional paragraphs of synthetic text, which began:

> The incident occurred on the downtown train line, which runs from Covington and Ashland stations.
>
> In an email to Ohio news outlets, the U.S. Department of Energy said it is working with the Federal Railroad Administration to find the thief.
>
> "The theft of this nuclear material will have significant negative consequences on public and environmental health, our workforce and the economy of our nation," said Tom Hicks, the U.S. Energy Secretary, in a statement. "Our top priority is to secure the theft and ensure it doesn't happen again."

Although this portion of the text was realistic, the model was far from perfect. OpenAI described several overarching failure modes of the model, including repetitive text, illogical combinations, and unnatural topic switching. In general, the company's researchers found that the model performed better on topics that were well represented in the training data. They also found the model did well at mimicking genres when provided with a specific subset of training data from a particular domain, such as the Amazon Reviews dataset.

Throughout the staged release process, parties outside of OpenAI experimented with the released models and found interesting additional uses. For example, a doctor at Imperial College in London retrained GPT-2 on a scientific database with over 30 million biomedical literature citations over a 24-hour period, and found the system could then create its own realistic and comprehensive scientific paper abstracts based merely on a title.[65] Others have used GPT-2 to generate poetry and write short stories.[66] A site called Talk to Transformer uses a simple interface to encourage people to experiment with the model using any custom prompt they choose.[67]

In addition to the staged release, OpenAI took other measures to support the goal of responsible publication. For example, along with releasing the code for GPT-2 on GitHub, OpenAI also published a "model card," which explains details about how the model was built and evaluated, includes research findings on biases in the model, provides an open communication channel for feedback, and gives recommended uses.[68] This idea was inspired by a paper originally published in late 2018 called "Model Cards for Model Reporting,"[69] which introduced the framework of model cards — short documents accompanying trained machine-learning models — to support greater transparency about a model's performance characteristics and other important information about how and why the model was built in a particular way.

The Final Release and Risk-Reduction Efforts

OpenAI released the largest version of GPT-2, with 1.5 billion parameters, on November 5, 2019.[70] In a blog accompanying the release, the team wrote, "While there have been larger language models released since August, we've continued with our original staged release plan in order to provide the community with a test case of a full staged release process. We hope that this test case will be useful to developers of future powerful models, and we're actively continuing the conversation with the AI community on responsible publication." In addition to the model, the company also released an updated report on social impacts,[71] as well as an updated model card.[72]

In the report, OpenAI's team provided insight into their process and findings from the staged release.[73] Part of this research took place in-house, including the release of research related to bias in the model's outputs. OpenAI also partnered with four outside organizations to focus on challenges of detection, misuse, and bias. Partner organizations included Cornell University; The Middlebury Institute of International Studies Center on Terrorism, Extremism, and Counterterrorism; The University of Oregon; and The University of Texas at Austin. These partnerships enabled investigation into potential malicious uses, detection of synthetic text, human responses to generated text, and biases in GPT-2 outputs. Justifying the company's cautious stance, the research found that people find GPT-2 outputs to be convincing, that the system can easily be fine-tuned for misuse, and that detection of synthetic text will be a long-term challenge. The OpenAI team also used the time between releases to engage with outside stakeholders, including by contributing to ongoing work carried out by the Partnership on AI on developing responsible publication norms.

As another precaution, the company communicated with outside researchers who were creating similar language models — a critical choice, as the practice of staged release works best in an environment of cooperation. OpenAI shared a specific GPT-2 email address and encouraged engagement from students and researchers. This feedback mechanism was reportedly used on numerous occasions. For example, the AI company Hugging Face decided against releasing its internal language models following discussion with OpenAI. The company describes itself as having a firm belief in open-source and knowledge sharing. Hugging Face has written, "Without open-source, the entire field faces the risk of not making progress and concentrating capabilities at the hands of a couple of massive players (be they corporations or states), without anyone else but them being able to understand, compete or control."[74] However, at the same time, the company acknowledges that its technology is not neutral and that action is required to facilitate its positive impact in the world, including considering the potential malicious uses of new releases. The company published an ethical analysis along with its latest conversational AI language model.[75] Similarly, when Salesforce released the language model CTRL, they also published an analysis discussing potential societal implications.[76]

OpenAI monitored uses of GPT-2 in the real world. They did this in part by tracking websites and forums with a history of promoting disinformation, as well as by having discussions with policymakers in defense and intelligence agencies. The team did not find significant evidence of misuse, though they acknowledge that advanced persistent threats (APTs) are particularly difficult to monitor. They also admitted to finding evidence of discussion of misuse, including a small number of cases of explicit public plans for misuse, though they did not believe the relevant actors had sufficient resources and capabilities to carry out the plans.

Reception

OpenAI has suggested that the experiment in staged release was relatively successful, mostly because it helped spur discussion about AI publication norms. Nonetheless, it was criticized for several reasons described below.

A key argument for transparency is that advanced language models can be used to support efforts to detect other fake media. For example, a research team at the University of

Washington released a different language model in June 2019, which they described as a "state-of-the-art generator of neural fake news."[77] These researchers disputed that releasing their model would be dangerous, arguing that the capabilities of GROVER and GPT-2 were not sufficiently human-like and that the lack of controllability of the content makes these models less useful for adversaries. Moreover, they argued that releasing the model in its entirety had benefits for threat modeling and defense. However, the GROVER researchers did discuss their work with people at OpenAI, and were encouraged to conduct in-depth threat modeling to inform their decision about how to release their model.

In November 2019, the cybersecurity company FireEye published a blog post revealing they were using GPT-2 to detect social media posts generated as part of information operations.[78] The company's researchers had fine-tuned the GPT-2 model on millions of tweets attributed to the Russian Internet Research Agency. This taught the model to create tweets that resembled the source, for example, "It's disgraceful that people have to waste time, energy to pay lip service to #Junk-Science #fakenews." However, the authors of the blog pointed out that, while they were able to use the system to detect malicious activity to spread propaganda, this model could also be used to lower the barrier to entry for actors hoping to engage in such malicious activity at scale.

Other criticisms of OpenAI's decision to delay the full release of GPT-2 stemmed from a perceived betrayal to core processes of peer review and the scientific method, as well as the culture of openness that has been central to AI progress for decades.[79] The belief in the value of open-source software to support replication and application has long been a central component of development in the field, key to avoiding another "winter" or period of diminished enthusiasm.[80] This critique was particularly sharp for OpenAI, which was founded on ideals of transparency and openness.[81] Critics of OpenAI's decision contended that the partial disclosure meant that independent researchers were not able to evaluate and verify claims made about the system, or to build upon previous findings. Some suggested that the company was overstating the uniqueness of its tool and engaging in fear-mongering.[82]

Indeed, in August 2019, two graduate students from Brown University announced they had replicated the 1.5 billion-parameter GPT-2 model by modifying the open-source Grover model.[83] Their purpose was to critique the strategy of staged released, which they argued only makes sense if a model is difficult to replicate. Instead, they proved that similar results could be recreated for roughly $50k by two masters students who had never created a language model before. The duo claimed that they were making a morally justified choice, writing, “Because our replication efforts are not unique, and large language models are the current most effective means of countering generated text, we believe releasing our model is a reasonable first step towards countering the potential future abuse of these kinds of models.”

Along with these critiques, OpenAI was also celebrated for the staged release decision, in particular for its impact on encouraging AI developers to think comprehensively about the implications of their work.[84] Norms within scientific communities can be powerful mechanisms to promote responsible innovation, and are particularly important in cutting-edge fields that have a relative lack of guidance and regulatory frameworks. OpenAI’s efforts have inspired other AI stakeholders — for example, the Partnership on AI — to consider the responsible publication of high-stakes AI research.[85]

Although there has been a long history of efforts to instill responsibility and ethics in technological developments, it is still rare for researchers and companies to offer transparent accounts of the risks stemming from their work. For example, while it is typical for AI researchers to state the positive uses of their models in papers, it is uncommon to see discussion about possible misuses or unintended consequences. Most technology companies are still wary to discuss the societal impact, risks, and potential negative implications of their products and services.

Today, there is less doubt about the risks that AI technologies pose. It has been well documented that AI systems that optimize for user engagement can promote extremism and filter bubbles,[86] that AI-enabled synthetic media (including deepfakes) can be used to generate malicious content,[87] and that AI systems can easily be tricked[88] and can make deadly mistakes.[89]

Impacts

Even if delaying the release of the largest GPT-2 model did little to prevent misuse of language models in general, OpenAI’s decision jump-started a larger conversation about best practices and responsible publication norms. The paper accompanying the release of the largest GPT-2 model concludes, “We hope GPT-2 as a case will help the AI community navigate publications in omni-use AI research.”[90] Their hope appears to have become reality, as others have subsequently adopted similar strategies.

For example, in January 2020, Google announced a new conversational agent called Meena, which integrates an astonishing 2.6 billion parameters. Meena is capable of engaging in conversations that are more realistic than current state-of-the-art systems. Importantly, Google decided not to release an external research demo of the system due to concerns about safety and bias, noting that the company is still evaluating the risks and benefits associated with giving the public access to this powerful tool.[91] Similarly, in November 2019, Microsoft announced a language model called DialoGPT, but did not include a public sampling interface in order to minimize opportunistic misuse.[92]

In September 2019, Salesforce released CTRL, a language model containing 1.63 billion parameters.[93] The researchers published their model in full and stated, “Openness and replicability are central aspects of the scientific ethos that, prima facie, suggest the release of complete scientific research results. We reify these principles by releasing all trained CTRL models.” However, their technical paper includes a section on “the ethics of large language models,” which reveals that they took responsible disclosure seriously. The researchers also published a second paper that delved more deeply into responsible innovation and the inadequacy of self-governance.[94]

The Salesforce research team was encouraged to engage on these issues because of the precedent set by others. Rather than rely on self-governance, they consulted with experts at the Partnership on AI, who have been working on the issue of responsible publication norms with members from OpenAI and other stakeholders. The Salesforce researchers carried out a technology foresight exercise that included scenario planning as a way to imagine worrisome possible uses of their technology. When Salesforce did release the CTRL model openly on GitHub, they included a code of conduct and a set of questions to “further encourage users to reflect on norms and responsibilities associated with models that generate artificial content.” Moreover, to facilitate post-release monitoring of CTRL, Salesforce actively observes how others are using CTRL. The team set up a dedicated email account and encourages users of the model to share their uses, pose questions, and suggest solutions.

Other organizations have opted for a more extreme stance on disclosing research results. The Machine Intelligence Research Institute (MIRI), whose mission is to ensure that the creation of smarter-than-human intelligence has a positive impact, released an update on its research directions in November 2018, which included a shift toward “nondisclosed-by-default research.”[95] The organization explained that the majority of its research results would no longer be published externally in an effort to prevent others from using their findings to build more capable and dangerous systems. This shift, which occurred three months before OpenAI's decision, was met with some confusion and skepticism, though OpenAI policy director Jack Clark called the move "useful" at the time, suggesting that it "generates data for the community about what the consequences are of taking such a step."[96]

AI researchers will continue to need to weigh the costs and benefits of different disclosure models. Some accountability measures, such as model cards, are likely to be beneficial in most cases, whereas appropriate degrees of openness will vary depending on the scale and scope of potential harm. Researchers at Oxford have proposed a theoretical framework to help inform this assessment. The framework addresses the security value of disclosure and includes factors that contribute to whether providing access to certain research will make it easier to cause harm, or easier to provide protections against harm. The factors include: counterfactual possession (i.e. where the would-be attacker acquires the relevant knowledge even without publication), absorption and application capacity (i.e. if publication of the research will only benefit attackers to the extent that they are able to absorb and apply the research), resources for solution-finding (i.e. given the disclosure, how many additional individuals or organizations will work on finding a solution?), availability of effective solutions (i.e. is there a good defense against misuse?), and the difficulty/cost of propagating a solution (i.e. even where a solution exists in theory, it might be difficult or costly to propagate that solution).[97]

This case study highlights how decisions about how to disclose omni-use AI research can have a lasting impact on the future of the field. Over the course of 2019, OpenAI undertook an experiment in responsible disclosure for advanced artificial intelligence by releasing progressively more capable versions of its powerful language model. The company used the time in between releases to monitor uses, engage with partner organizations on particular research questions, and promote awareness of impacts. This decision has already had a significant impact on other AI researchers and organizations, and is poised to have an outsized effect on future trajectories of AI development. It is becoming more normal to see open deliberations about risks and tradeoffs inherent to AI systems, even as new models are made publicly available around the world. Keeping research as open as possible, while minimizing the potential for misuse and harm, is a delicate balance.

CASE STUDY III

CAN A GLOBAL FOCAL POINT FOR AI POLICY ANALYSIS AND IMPLEMENTATION PROMOTE INTERNATIONAL COORDINATION?

The Launch of the OECD AI Policy Observatory

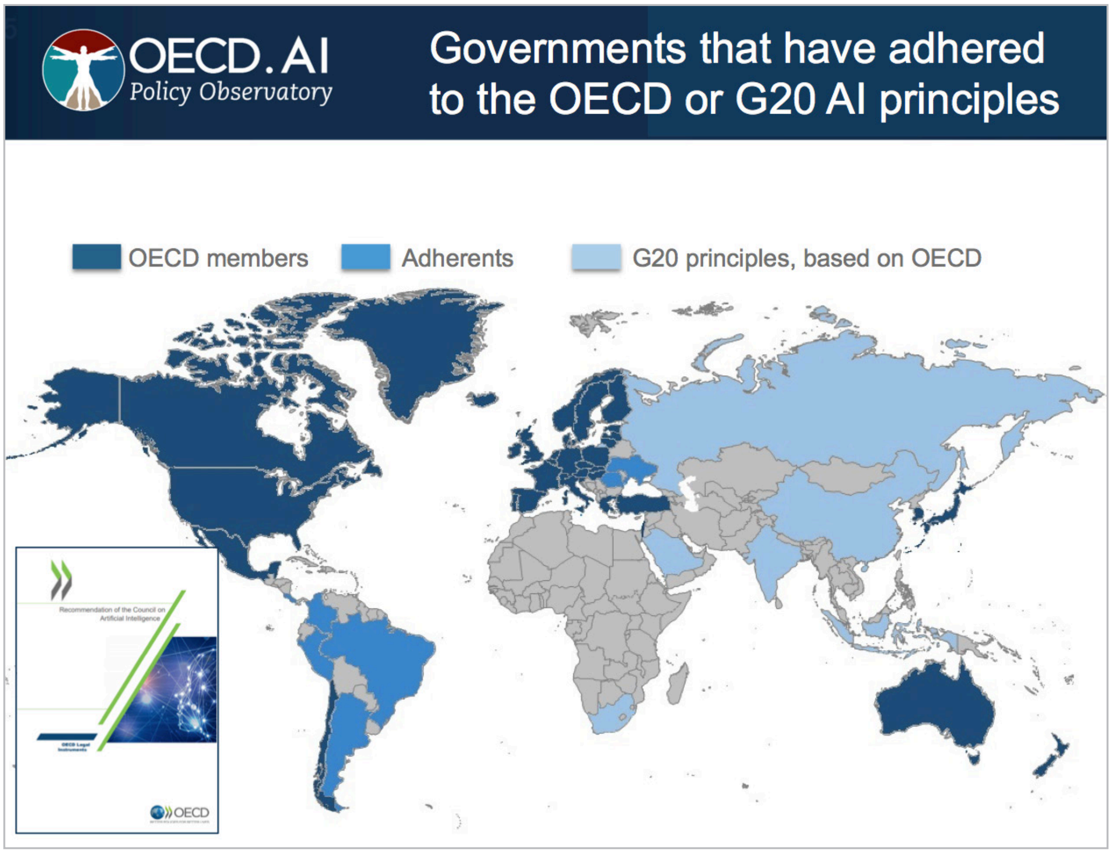
INSIGHTS

- International coordination and cooperation on AI begins with a common understanding of what is at stake and what outcomes are desired for the future. That shared language now exists in the Organisation for Economic Co-operation and Development (OECD) AI Principles, which are being leveraged to support partnerships, multilateral agreements, and the global deployment of AI systems.
- Stakeholders largely agree on high-level interests such as AI safety and transparency, but there will continue to be differences in the implementation of AI principles within different political and economic environments.
- Evidence-based AI policy guidance, metrics, and case studies to support domestic AI policy decisions are in high demand, and the OECD AI Policy Observatory is poised to become a prominent source of guidance globally.
- The function of the OECD AI Policy Observatory as an intergovernmental hub for AI governance may serve as a counterpoint to AI nationalism and the prominence of "AI race" rhetoric between nations.

The First Intergovernmental Standard for AI

It has become common to hear about the "race for AI supremacy" between nations,[98] despite known dangers associated with such rhetoric.[99] In particular, the focus on national advantage can undercut efforts to support a global approach to AI governance. This may be problematic for AI technologies, which are becoming widely available around the world, and whose effects will be far-reaching. Given the dynamics around national competition, it would have been difficult to predict that dozens of nations, and especially the U.S., China, and Russia, would agree to a common set of guiding principles for AI. However, that is what happened in June 2019, when the G20 gave unanimous support to the OECD AI Principles. This case study explores the events that led up to that occasion (and what was left out of the agreement), the policy mechanisms planned to support the implementation of the principles around the world, and what the developments mean for the global governance of AI. This analysis is based upon interviews and feedback from OECD employees and expert group members, and a review of public documentation and media.

On May 22, 2019, forty-two countries adopted the first intergovernmental standard on artificial intelligence.[100] The guidelines came in the form of a legal recommendation that included five principles and five recommendations from the OECD, led by the Committee on Digital Economy Policy (CDEP). In announcing this initiative, OECD Secretary-General Angel Gurría stated, "These Principles will be a global reference point for trustworthy AI so that we can harness its


OECD.AI Policy Observatory

Governments that have adhered to the OECD or G20 AI principles

OECD members
Adherents
G20 principles, based on OECD
Recommendation of the Council on Artificial Intelligence
OECD Legal Instruments
OECD

Image used with permission from a presentation by Karine Perset, Administrator on Digital Economy and Artificial Intelligence Policy in the OECD Directorate for Science, Technology, and Innovation, and by Adam Murray, International Affairs Officer in the U.S. Department of State Office of International Communications and Information Policy, delivered to members of the Partnership on AI in April 2020. The map shows that there has been a broad global commitment to the OECD AI Principles, but also highlights that many African nations have not yet been involved.

opportunities in a way that delivers the best outcomes for all." Unlike other sets of AI principles, the OECD AI Principles are an intergovernmental agreement; although the process to develop them brought together multiple stakeholders, the adherents are governments, making this the first intergovernmental standard for AI in existence.

All 36 OECD member countries signed on to the OECD AI Principles, including many nations at the forefront of AI development, among them the United States, Australia, France, Germany, Korea, Estonia, Israel, Japan, and the United Kingdom. Several non-member countries — including Argentina, Brazil, Colombia, Costa Rica, Peru and Romania — also signed on. The European Commission additionally supported the Principles, and Ukraine was added to the list of signatories in October 2019. When the Group of Twenty (G20) released AI Principles one month later, it was noted that they were drawn from the OECD AI Principles,[101] expanding the list of supportive countries to include China, India, and Russia, among others.

Established in 1961 as an intergovernmental organization, the Paris-based OECD today has 36 member countries in Europe, North America, South America, and Asia. All of the members are market-based democracies, and the organization has been criticized for being a "club of mostly rich countries."[102] However, the OECD has been expanding its membership over the years, and has started to partner with more developing countries, including Brazil, India, and South Africa.[103] The OECD focuses on economic and social policy analysis and statistics, and also develops international policy standards, such as the OECD Privacy Guidelines and the OECD AI Principles.

The OECD has made 177 policy recommendations since 1964, and has a history of promoting international cooperation on the safety of consequential, dual-use technologies, including genetic engineering and nuclear technology. However, that the organization would play such a prominent role in the global governance of AI was not a given. Other institutions, such as the United Nations and International Telecommunication Union (ITU), have also emerged as forums for advancing the global governance of AI, but have not similarly garnered support for ethics and governance principles to date.[104] The ability of the OECD AI Principles to attract the support of dozens of governments was several years in the making.

The OECD's Committee on Digital Economy Policy (CDEP) began considering a recommendation on AI as early as 2016, and in May 2018, this committee decided to establish an AI expert group to scope AI principles. The expert group (AIGO) launched in September 2018 with 50 members, led by Wonki Min, Vice Minister of Science and ICT of Korea and chair of the OECD's Digital Economy Committee. Many countries were represented among the AIGO members, including: Australia, Canada, Denmark, Finland, France, Germany, Hungary, Japan, Korea, Mexico, Netherlands, New Zealand, Poland, Russia, Singapore, Slovenia, Sweden, Switzerland, Turkey, UAE, United Kingdom, United States, and the European Commission.

In addition to government representatives, the group invited experts from industry, academia, and civil society, including from Microsoft, Google, Facebook, MIT, the Harvard Berkman Klein Center, OpenAI, IEEE, the AI Initiative of The Future Society, and the World Privacy Forum, among others. Contributions from other experts around the world were also taken into account. Broad, multistakeholder engagement and significant enthusiasm and dedication from group members were pivotal for the OECD's success in advancing a global governance framework for AI. The group helped to scope the principles over four in-person meetings in different locations around the world (Paris, Cambridge, and Dubai), and with several teleconference calls in between. This group initially identified the five principles and recommendations, which were then expanded upon further.[105]

Some governments played a more integral role in enabling the OECD to meaningfully tackle the AI governance challenge. For example, the OECD's work on AI began in April 2016 at the G7 ICT Ministerial meeting in Takamatsu, Japan, where the host nation encouraged the OECD to prioritize AI and identify policy priorities for international cooperation. Japanese ministers described the need for international principles to guide research and development of AI, and proposed an initial set of principles for consideration that included transparency, user assistance, controllability, security, safety, privacy, ethics, and accountability. Japan's Ministry of Internal Affairs and Communications (MIC) also provided financial support for an OECD conference, "AI: Intelligent Machines, Smart Policies," a landmark event held in Paris in October 2017. [106] The MIC also supported the development of the book *Artificial Intelligence in Society*, published by the OECD in June 2019, which provides greater background about the emergence of the principles and describes policy initiatives under way around the world.

Moreover, Japan proposed and led the G20 discussion on AI, facilitating the agreement to the G20 AI Principles at the Ministerial Meeting on Trade and Digital Economy in Osaka, Japan. When Japanese Prime Minister Shinzo Abe announced the G20 AI Principles, he confirmed that they would guide the G20's commitment to a human-centered approach to AI. Japan's ongoing leadership and support for a global AI governance framework has been critically important. Interviewees have confirmed that if Japan had not held the G20 presidency in 2019, the G20 AI

Principles would not exist. Japan will continue to be a key stakeholder in the global governance of AI, and has indicated its intention of continuing to support the OECD's efforts.

The United States government has also been a vocal supporter of the OECD AI Principles. In a speech at the OECD forum and ministerial council meeting in Paris, Michael Kratsios, then Deputy Assistant to the President for Technology Policy (and now Chief Technology Officer), referred to the moment as "a historic step," by which, "America and likeminded democracies of the world will commit to common AI principles reflecting our shared values and priorities."[107] Kratsios commented, "The United States has long welcomed the work of the OECD to develop AI principles. Across multiple G7 and OECD fora, we worked closely with our strong international partners to advance discussions and draft the principles." The United States in particular has focused on the importance of identifying the shared values of democratic nations for the development of AI.

G20 Support

Held on June 28, 2019, the G20 Osaka Summit brought together G20 leaders in Japan to address major global economic challenges. Invited international organizations included the United Nations and the World Bank. This annual meeting, primarily intended to coordinate responses to global economic turbulence, had an increased focus on the role of digitalization and technological innovation.

Nonetheless, at the outset of the Summit, it was not widely expected that the group would reach an agreement on guidelines related to artificial intelligence. Yet, by the conclusion, the group released the G20 AI Principles,[108] which were accepted by consensus and established a common set of principles for the responsible stewardship of trustworthy AI. The development of a common set of principles for AI development among nations with diverse and at times conflicting interests, including the U.S. and China, was a shocking and important achievement. Within a culture of national competition for AI leadership, the G20 Principles represented a first step toward collective action on AI governance at the global scale.

A footnote in the G20 AI Principles notes that that they were drawn from the OECD principles and recommendations for artificial intelligence. However, the G20 AI Principles are in fact largely identical to the OECD AI Principles. Both documents include the following language:

1. **Inclusive growth, sustainable development and well-being**
   Stakeholders should proactively engage in responsible stewardship of trustworthy AI in pursuit of beneficial outcomes for people and the planet, such as augmenting human capabilities and enhancing creativity, advancing inclusion of underrepresented populations, reducing economic, social, gender and other inequalities, and protecting natural environments, thus invigorating inclusive growth, sustainable development and well-being.
2. **Human-centered values and fairness**
   a. AI actors should respect the rule of law, human rights and democratic values, throughout the AI system lifecycle. These include freedom, dignity and autonomy, privacy and data protection, non-discrimination and equality, diversity, fairness, social justice, and internationally recognized labor rights.
   b. To this end, AI actors should implement mechanisms and safeguards, such as capacity for human determination, that are appropriate to the context and consistent with the state of art.
3. **Transparency and explainability**
   AI Actors should commit to transparency and responsible disclosure regarding AI systems. To this end, they should provide meaningful information, appropriate to the context, and consistent with the state of art:
   a. to foster a general understanding of AI systems;
   b. to make stakeholders aware of their interactions with AI systems, including in the workplace;
   c. to enable those affected by an AI system to understand the outcome; and,
   d. to enable those adversely affected by an AI system to challenge its outcome based on plain and easy-to-understand information on the factors, and the logic that served as the basis for the prediction, recommendation or decision.
4. **Robustness, security and safety**
   a. AI systems should be robust, secure and safe throughout their entire lifecycle so that, in conditions of normal use, foreseeable use or misuse, or other adverse conditions, they function appropriately and do not pose unreasonable safety risk.
   b. To this end, AI actors should ensure traceability, including in relation to datasets, processes and decisions made during the AI system lifecycle, to enable analysis of the AI system's outcomes and responses to inquiry, appropriate to the context and consistent with the state of art.
   c. AI actors should, based on their roles, the context, and their ability to act, apply a systematic risk management approach to each phase of the AI system lifecycle on a continuous basis to address risks related to AI systems, including privacy, digital security, safety and bias.

5. **Accountability**
   AI actors should be accountable for the proper functioning of AI systems and for the respect of the above principles, based on their roles, the context, and consistent with the state of art.

G20 support for the OECD AI Principles was particularly meaningful for several reasons. The G20 countries account for about 85% of global economic output, 75% of global exports, and two-thirds of the world's population.[109] The G20 has become a premier forum for international cooperation and coordination, and consensus support from G20 countries expands the reach of the AI principles around the world. Although the primary focus of the G20 is the global economy, recent meetings have increasingly been used to discuss pressing foreign policy challenges, ranging from sustainability to human rights abuses.

However, G20 support did not extend to the full content of the OECD AI Recommendation. The second section of the recommendation, "National policies and international co-operation for trustworthy AI," reveals some differences of opinion among the OECD and G20 countries. The recommended policies include (at a high level) investing in AI research and development, fostering a digital ecosystem for AI, shaping an enabling policy environment for AI, building human capacity and preparing for labor market transformation, and international co-operation for trustworthy AI. An appendix to the "G20 Ministerial Statement on Trade and Digital Economy" states, "The G20 supports the Principles for responsible stewardship of Trustworthy AI in Section 1 and takes note of the Recommendations in Section 2."[110] In other words, while the G20 endorsed the OECD AI Principles, the support did not explicitly extend to the recommendations for governments.

This fact underscores the need to track and analyze the operationalization of AI principles and strategies globally. While the development of intergovernmental principles for AI was remarkable, how these goals are realized is not likely to be uniform. This gap is one of the most common criticisms of the OECD AI Principles: that they are too high-level to lead to real policy change. This is a relevant critique of all voluntary principles, and should not be taken lightly. Like all other G20 declarations, the G20 AI Principles are non-binding and their full impact remains to be seen. The G20 has been criticized for not doing more than "naming and shaming" when actors fail to uphold their commitments. The OECD Recommendation is also not legally binding, and the OECD lacks enforcement capabilities. Nonetheless, other OECD Recommendations have been quite influential as political commitments, in particular for setting international standards and helping with the design of national legislation. For example, the OECD Privacy

Guidelines influenced the design of privacy laws around the world. In this case, the OECD developed an additional plan to support the practical and policy relevance of its principles.

The AI Policy Observatory

At the end of 2019, the OECD announced plans for an AI Policy Observatory to help countries implement the principles and recommendations.[111] Formally launched in late February 2020, the Observatory is envisioned as "a platform to share and shape public policies for responsible, trustworthy and beneficial AI." The Observatory supports dialogue among global multistakeholder partners, publishes practical guidance to implement the AI Principles, and supports a live database of AI policies and initiatives globally. It also compiles metrics and measurement of AI development to serve as a baseline for policy development, and uses its convening power to bring together the private sector, governments, academia, and civil society. The Observatory's resources have all been made publicly available at OECD.ai.

The Observatory is structured around four main pillars, each of which has its own goals, partners, and online resources. The first pillar centers on the OECD AI Principles. The objectives of the pillar include explaining what each principle means and why it matters, and providing practical guidance to governments, including resources to support implementation efforts. Dashboards dedicated to each principle provide concrete information about related AI policy initiatives, policy instruments, and scientific research. This is considered to be a key goal of the Observatory. As Karine Perset, Administrator on Digital Economy and Artificial Intelligence Policy in the OECD Directorate for Science, Technology, and Innovation, emphasized, "The principles were the beginning, and now we are focusing on implementation. The AI Policy Observatory is one of our major endeavors to move from principles to action and implementation, and help policymakers in this journey."[112]

As the second pillar, the Observatory will provide analysis of AI policy in key areas, including science, health, jobs, and transportation, among others. Dashboards dedicated to each sector provide information about related policy initiatives, live updated news feeds, and how different countries are prioritizing research in that area. The third pillar is focused on trends and data, and includes OECD metrics and measurement, as well as live data from partners, including news about AI development and policy. For example, you can explore data and visualizations that depict trends in publications from different countries, AI research collaborations and networks, the growth of AI subtopics, and AI skills migration between countries. These resources are intended to help provide a basis for evidence-based policymaking.

The fourth pillar relates to national AI strategies, policies, and initiatives, from national governments and other AI stakeholders. This includes a database, visualizations, and analysis of over 300 AI policy examples, which are contributed directly from governments through a survey. This resource serves as a unique repository that enables countries and organizations to compare AI policies at a much more granular level than has previously been possible. For example, the dashboards show all of the relevant initiatives under way in a given country, the prioritization and investment amounts in different areas, and relevant governmental bodies and research institutions.

Numerous global data streams are used to inform the insights on OECD.ai, including news media, scientific papers, patents, job market data, and business data. The Observatory is also informed by the Microsoft Academic Graph, which collects information about scientific publications, citation relationships, authors, institutions, journals, conferences, and fields of study; and the LinkedIn Economic Graph, which uses the company's data to highlight trends related to talent migration, hiring rates, and in-demand skills by region. The Observatory makes use of AI techniques, including social network analysis and classification, to process and analyze this data before presenting it in a relatively interactive and accessible way. The data is intended not only to provide a view of the past and present, but also to provide views of future trends. Global policymakers are the primary audience of the resource, and it is hoped that they can utilize the insights to inform evidence-based practices and policy development.

The OECD AI Policy Observatory already has support from numerous governments and multistakeholder organizations. For example, UNESCO (The United Nations Educational, Scientific, and Cultural Organization) aims to support the OECD's AI governance priorities by translating policy recommendations into actionable opportunities for the communities they work with in their field offices around the world. The European Commission also intends to support the Observatory, especially in its work on metrics and measurement and the collection of national AI strategies and policies. The governments of Germany, Japan, and the United States have continued to publicly voice their support of the Observatory and seek alignment between their national AI policy initiatives and the OECD AI Principles.[113] Another emerging international initiative called the Global Partnership for AI (GPAI), led by French president Emmanuel Macron and Canadian prime minister Justin Trudeau, will coordinate with the OECD to provide a forum for global debate on AI.[114]

A new OECD Network of Experts on AI (ONE AI) will also advise and support the work of the AI Policy Observatory (replacing and building upon the former expert group, AIGO).[115] ONE AI is a multi-stakeholder and multi-disciplinary advisory group that is composed of more than 100 AI experts split into three working groups. The groups provide input and implementation ideas for AI policy issues, support the Observatory's four pillars, and facilitate information exchange and collaboration between the OECD and other international initiatives and organizations focusing on AI. For example, at the group's first meeting on February 27, 2020, discussions centered on classifications of AI systems and ongoing efforts to implement practices that support safe and human-centric AI.

It was recognized at that meeting that governments will need to adopt new policies and practices across numerous sectors to ensure the principles and recommendations are adopted. There is some indication that this is already happening. For example, the European Commission, which has adopted the OECD AI Principles, is developing legislative proposals for AI and intends to facilitate a coordinated European approach to the human and ethical implications of AI.[116] Moreover, the OECD AI Principles are referenced as a meaningful source for AI standards in the August 2019 Plan for Federal Engagement in Developing Technical Standards and Related Tools, developed by the US Department of Commerce National Institute of Standards and Technology (NIST).[117]

G20 support for the OECD AI Principles and Observatory is also likely to continue. In his concluding declaration before the other global leaders, Japanese Prime Minister Shinzo Abe expressed the shared commitment to "human-centered" AI and suggested that AI is a "driving force" behind the sustainable development goals (SDGs). Prime Minister Abe acknowledged "the growing importance of promoting security in the digital economy and of addressing security gaps and vulnerabilities." He also discussed more broadly the increasing importance of digital technologies and the cross-border flow of data for innovation and the global economy. Saudi Arabia assumed the G20 Presidency in 2020 and has clarified its goals of supporting the development of inclusive and trustworthy AI.[118] The implementation of AI principles within different sectors was a focus at the first G20 Digital Economy Task Force meeting in February 2020 in Riyadh, Saudi Arabia. Interviewees involved with the G20 suggested that digitalization and AI are likely to be on the agenda for many years to come.

Despite the hurdles ahead, the decision to support a common set of AI principles marked a critical shift in the global landscape. The OECD AI Principles represent the first time that nations around the world committed to a common set of guidelines that provide shared understanding and goals for how to shape future trajectories of AI. The OECD Principles are also notable compared to other AI principles for highlighting a broader range of issues, including reducing economic, social, gender and other inequalities, protecting natural environments and internationally recognized labor rights, and applying a systematic risk management approach to each phase of the AI system lifecycle on a continuous basis.

Perhaps most importantly, the OECD AI Policy Observatory will, by design, ensure that the Principles are linked to concrete policy mechanisms that can be implemented by nations around the world. This will help to operationalize the AI Principles at a large scale. Moreover, the Observatory is poised to become a prominent site for multistakeholder dialogue on AI. The Observatory's openly available online materials and regular events will facilitate international coordination and collaboration at a scale that has previously been difficult to sustain. Other initiatives at the OECD, such as the OECD Global Parliamentary Network — a learning hub on AI for legislators and parliamentary officials — will help ensure a comprehensive approach to achieving the organization's goals.

The OECD AI Principles achieved a feat few would have thought possible. The United States signed on at a time of relative aversion to international coordination in other policy arenas. China was part of a consensus agreement to support the effort more broadly, and other countries are welcome to add their support. The year 2019 brought the first intergovernmental standard for AI and a new "global reference point" for AI governance into the future. Moreover, the OECD AI Policy Observatory has been identified by many of the world's governments as the new global focal point for translating AI principles into practical policy guidelines. Sustained attention to the coming successes and challenges of these efforts will be important to further understand the value of this model. At this point, the AI Policy Observatory is the first initiative of its kind and is poised to have an outsized impact on the trajectory of AI around the world.

Conclusion

AI stakeholders face countless difficult decisions about how, why, and for whom to develop and use AI technologies. The concept of "AI decision points" provides a framework to prioritize decisions that were not predetermined by existing law or practice, and that mark a meaningful shift in behavior from previous practice for shaping the development and use of AI. These decisions are catalysts for broader inflection points and are reshaping future trajectories of AI. Identifying and tracking AI decision points can provide insight into the evolution of the field, and help focus governance efforts by identifying key policy levers.

Decisions about how to operationalize AI principles and strategies are currently faced by nearly all AI stakeholders, and are determining practices and policies in a meaningful way. There is growing pressure on AI companies and organizations to adopt implementation efforts, and those actors perceived to verge from their stated intentions may face backlash from employees, users, and the general public. The transition from principles to practice in the AI field has become shorthand for the broad desire to see plans put into action, through organizational shifts, design decisions, or the implementation of new policies. Nonetheless, this is understood to be a difficult and time-consuming process: legal frameworks are shifting, and no taxonomies of best practices have been agreed upon.

This paper aims to provide a step in that direction, by compiling examples of efforts to bridge the gap between principles and practice, and by focusing on case studies that are meaningful examples of this translation process. The case studies reveal insights about how companies and intergovernmental institutions are approaching decisions about the operationalization of AI principles, as well as the challenges and successes each effort has faced.

The example of Microsoft's AETHER Committee highlights the importance of executive-level support in shaping an organization's commitment to responsible AI development, as well as the value of employee and expert engagement, and integration with the company's legal team. AETHER's structure has enabled the establishment of new organizational policies and the prioritization of new areas of research. This model would be less valuable if internal dynamics at Microsoft were to diminish AETHER's decision-making power, or if the regulatory landscape were to shift to reduce individual companies' ability to make decisions about the design and sale of AI technologies.

The review of OpenAI's staged release of the GPT-2 AI language model highlights the spectrum between "open" and "closed" AI research, as well as the difficulties of preventing consequential technologies from being misused. This case study exemplifies how companies can make use of multiple synergistic accountability measures, including documentation efforts, discussion of potentially harmful uses and impacts in research papers, and facilitating communication prior to and following the release of new AI models.

Finally, the examination of the OECD AI Policy Observatory highlights how, despite challenges in achieving international cooperation, governments remain motivated to support global governance frameworks for AI. While the Observatory is still in its infancy, governments, like companies, are seeking guidance on actions they can take to realize their objectives for responsible AI. Though it may one day be superseded by other intergovernmental forums or treaties, the Observatory has emerged as an important resource for nations to share evidence-based AI policy guidance and metrics, and to facilitate global dialogue.

Together, the case studies shine a light on how influential AI stakeholders are navigating the "third stage" of AI governance, translating principles into practice. Given implementation efforts are dependent on context, case studies and ethnographic accounts can help illuminate how the field is shifting to address concerns about the significant safety, security, and societal challenges accompanying the evolution of artificial intelligence. Decisions made today about how to operationalize AI principles at scale will have major implications for decades to come, and AI stakeholders have an opportunity to learn from existing efforts and to take concrete steps to ensure we build a better future.

Endnotes

[1] Raymond Perrault, Yoav Shoham, Erik Brynjolfsson, Jack Clark, John Etchemendy, Barbara Grosz, Terah Lyons, James Manyika, Saurabh Mishra, and Juan Carlos Niebles, "The AI Index 2019 Annual Report," AI Index Steering Committee, Human-Centered AI Institute, Stanford University, December 2019, https://hai.stanford.edu/sites/g/files/sbiybj10986/f/ai_index_2019_report.pdf.

[2] Anna Jobin, Marcello Ienca, and Effy Vayena, "Artificial Intelligence: the global landscape of ethics guidelines," *Nature Machine Intelligence* volume 1, September 2019, https://www.nature.com/articles/s42256-019-0088-2.

[3] Jessica Fjeld, Nele Achten, Hannah Hilligoss, Adam Nagy, and Madhulika Srikumar, "Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI," Berkman Klein Center Research Publication No. 2020-1, January 15, 2020, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3518482.

[4] Jessica Cussins Newman, "Toward AI Security: Global Aspirations for a More Resilient Future," CLTC White Paper Series, February 2019, https://cltc.berkeley.edu/wp-content/uploads/2019/02/CLTC_Cussins_Toward_AI_Security.pdf.

[5] Jessica Fjeld, Nele Achten, Hannah Hilligoss, Adam Nagy, and Madhulika Srikumar, "Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI," Berkman Klein Center Research Publication No. 2020-1, January 15, 2020, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3518482.

[6] Jessica Morley, Luciano Floridi, Libby Kinsey, and Anat Elhalal, "From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices," *Science and Engineering Ethics*, December 11, 2019, https://link.springer.com/article/10.1007/s11948-019-00165-5.

[7] Svetlana Kiritchenko and Saif M. Mohammad, "Biases in Sentiment Analysis Systems," May 2018, https://saifmohammad.com/WebPages/Biases-SA.html.

[8] "InterpretML," *GitHub*, https://github.com/interpretml/interpret.

[9] "Welcome to the Adversarial Robustness 360 Toolbox," IBM Corporation, 2018, https://adversarial-robustness-toolbox.readthedocs.io/en/latest/.

[10] Kush R. Varshney, "Introducing AI Fairness 360," IBM Research Blog, September 19, 2018, https://www.ibm.com/blogs/research/2018/09/ai-fairness-360/.

[11] Carey Radebaugh and Ulfar Erlingsson, "Introducing TensorFlow Privacy: Learning with Differential Privacy for Training Data," TensorFlow Blog, March 06, 2019, https://blog.tensorflow.org/2019/03/introducing-tensorflow-privacy-learning.html.

[12] Adele Peters, "This tool lets you see—and correct—the bias in an algorithm," *Fast Company*, June 12, 2018, https://www.fastcompany.com/40583554/this-tool-lets-you-see-and-correct-the-bias-in-an-algorithm.

[13] Melissa Mulholland, "Our shared responsibility for AI," Microsoft Blog, November 7, 2018, https://blogs.partner.microsoft.com/mpn/shared-responsibility-ai-2/.

[14] Kent Walker, “Google AI Principles updates, six months in,” The Keyword, December 18, 2018, https://www.blog.google/technology/ai/google-ai-principles-updates-six-months/.

[15] “DeepMind Ethics & Society Fellows,” DeepMind, 2019, https://deepmind.com/about/ethics-and-society#fellows.

[16] “Axon AI and Policing Technology Ethics Board,” Axon, April 2018, https://www.axon.com/axon-ai-and-policing-technology-ethics.

[17] Margaret Mitchell, et al., “Model Cards for Model Reporting,” *arXiv*, October 5, 2018, https://arxiv.org/abs/1810.03993.

[18] Timnit Gebru et al., “Datasheets for Datasets,” *arXiv*, January 15, 2020, https://arxiv.org/pdf/1803.09010.pdf.

[19] “GPT-2: 1.5B Release,” OpenAI, November 5, 2019, https://openai.com/blog/gpt-2-1-5b-release/.

[20] Inioluwa Deborah Raji et al., “Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing,” *arXiv*, January 3, 2020, https://arxiv.org/pdf/2001.00973.pdf.

[21] “Ethics guidelines for trustworthy AI,” European Commission, April 8, 2019, https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

[22] “The Trust-in-AI Framework,” a3i, 2019, http://a3i.ai/trust-in-ai.

[23] “The AI-RFX Procurement Framework,” The Institute for Ethical AI & ML, 2019, https://ethical.institute/rfx.html.

[24] “Algorithmic Accountability Policy Toolkit,” AI Now Institute, October 2018, https://ainowinstitute.org/aap-toolkit.pdf.

[25] Amanda Russo, “Artificial Intelligence Toolkit Helps Companies Protect Society and Their Business,” World Economic Forum, January 17, 2020, https://www.weforum.org/press/2020/01/artificial-intelligence-toolkit-helps-companies-protect-society-and-their-business/.

[26] Dillon Reisman et al., “Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability,” AI Now Institute, April 2018, https://ainowinstitute.org/aiareport2018.pdf.

[27] “Human rights impact assessments,” Business & Human Rights Resource Centre, 2019, https://www.business-humanrights.org/en/un-guiding-principles/implementation-tools-examples/implementation-by-companies/type-of-step-taken/human-rights-impact-assessments.

[28] Bin Yu and Karl Kumbier, “Veridical Data Science,” *arXiv*, January 23, 2019, https://arxiv.org/abs/1901.08152.

[29] “ISO/IEC JTC 1 Information Technology,” ISO, 2019, https://www.iso.org/isoiec-jtc-1.html.

[30] “The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems,” IEEE SA, https://standards.ieee.org/industry-connections/ec/autonomous-systems.html.

[31] For examples, see “From Principles to Practice: Ethically Aligned Design Conceptual Framework” from IEEE; “Bridging AI Principles to Practice with ABOUT ML” from the Partnership on AI; and the launch video of the OECD AI Policy Observatory.

[32] Sundar Pichai, "AI at Google: our principles," The Keyword, Google, June 7, 2018, https://www.blog.google/technology/ai/ai-principles/.

[33] Alexia Fernández Campbell, "The employee backlash over Google's censored search engine for China, explained," *Vox*, August 17, 2018, https://www.vox.com/2018/8/17/17704526/google-dragonfly-censored-search-engine-china.

[34] Thomas Brewster, "Microsoft Slammed For Investment In Israeli Facial Recognition 'Spying On Palestinians'," *Forbes*, August 1, 2019, https://www.forbes.com/sites/thomasbrewster/2019/08/01/microsoft-slammed-for-investing-in-israeli-facial-recognition-spying-on-palestinians/.

[35] Satya Nadella, "Satya Nadella email to employees: Embracing our future: Intelligent Cloud and Intelligent Edge," Microsoft News Center, March 29, 2018, https://news.microsoft.com/2018/03/29/satya-nadella-email-to-employees-embracing-our-future-intelligent-cloud-and-intelligent-edge/.

[36] Eric Horvitz, "Advancing Human-Centered AI," Microsoft Research Blog, March 18, 2019, https://www.microsoft.com/en-us/research/blog/advancing-human-centered-ai/.

[37] Karen Hao, "In 2020, let's stop AI ethics-washing and actually do something," *MIT Technology Review*, December 27, 2019, https://www.technologyreview.com/s/614992/ai-ethics-washing-time-to-act/.

[38] Brad Smith and Harry Shum, *The Future Computed: Artificial Intelligence and its role in society*, Microsoft, 2018, https://blogs.microsoft.com/uploads/2018/02/The-Future-Computed_2.8.18.pdf.

[39] Nick Statt, "Google dissolves AI ethics board just one week after forming it," *The Verge*, April 4, 2019, https://www.theverge.com/2019/4/4/18296113/google-ai-ethics-board-ends-controversy-kay-coles-james-heritage-foundation.

[40] Eric Horvitz, "Keynote Address, Eric Horvitz: AI Advances, Aspirations—and Concerns," Bulletin of the Atomic Scientists, November 15, 2019, https://www.youtube.com/watch?v=TUtUTZvZ1-4.

[41] Andrew Marshall, Jugal Parikh, Emre Kiciman and Ram Shankar Siva Kumar, "Threat Modeling AI/ML Systems and Dependencies," Microsoft, November 10, 2019, https://docs.microsoft.com/en-us/security/threat-modeling-aiml.

[42] Samuel Jenkins, Harsha Nori, Paul Koch, and Rich Caruana, "InterpretML - Alpha Release," *GitHub*, 2019, https://github.com/interpretml/interpret.

[43] Alan Boyle, "Microsoft is turning down some sales over AI ethics, top researcher Eric Horvitz says," *GeekWire*, April 9, 2018, https://www.geekwire.com/2018/microsoft-cutting-off-sales-ai-ethics-top-researcher-eric-horvitz-says/.

[44] Joseph Menn, "Microsoft turned down facial-recognition sales on human rights concerns," *Reuters*, April 16, 2019, https://www.reuters.com/article/us-microsoft-ai-idUSKCN1RS2FV.

[45] Olivia Solon, "Microsoft Funded Firm Doing Secret Israeli Surveillance on West Bank," *NBC News*, October 28, 2019, https://www.nbcnews.com/news/all/why-did-microsoft-fund-israeli-firm-surveils-west-bank-palestinians-n 1072116.

[46] Olivia Solon, "MSFT Hires Eric Holder to Audit AnyVision's Facial Recognition Tech," *CNBC*, November 15, 2019, https://www.cnbc.com/2019/11/15/msft-hires-eric-holder-to-audit-anyvisions-facial-recognition-tech.html.

[47] Brad Smith, "Technology and the US military," Microsoft Blog, October 26, 2018, https://blogs.microsoft.com/on-the-issues/2018/10/26/technology-and-the-us-military/.

48 “Keynote Address, Eric Horvitz: AI Advances, Aspirations—and Concerns,” Bulletin of the Atomic Scientists, YouTube, November 15, 2019, https://www.youtube.com/watch?v=TUtUTZvZ1-4&feature=youtu.be.

49 Microsoft Workers 4 Good, Twitter, February 22, 2019, https://twitter.com/MsWorkers4/status/1099066343523930112.

50 Michael Madaio and Jennifer Wortman Vaughan, “Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI,” 2020, https://www.semanticscholar.org/paper/Co-Designing-Checklists-to-Understand-Challenges-in-Madaio-Vaughan/58bb221c1e375f254826b7b7341f74057e87676c.

51 Natalia Drozdiak, “Microsoft Seeks to Restrict Abuse of its Facial Recognition AI,” *Bloomberg*, January 22, 2019, https://www.bloomberg.com/news/articles/2019-01-23/microsoft-seeks-to-restrict-abuse-of-its-facial-recognition-ai.

52 Julie Brill, “Our support for meaningful privacy protection through the Washington Privacy Act,” Microsoft On the Issues, April 29, 2019, https://blogs.microsoft.com/on-the-issues/2019/04/29/our-support-for-meaningful-privacy-protection-through-the-washington-privacy-act/.

53 “Microsoft expands artificial intelligence (AI) efforts with creation of new Microsoft AI and Research Group,” Microsoft News Center, September 29, 2016, https://news.microsoft.com/2016/09/29/microsoft-expands-artificial-intelligence-ai-efforts-with-creation-of-new-microsoft-ai-and-research-group/.

54 Martin Giles, “Microsoft is launching a huge reorganization to focus on AI and the cloud,” *MIT Technology Review*, March 29, 2018, https://www.technologyreview.com/f/610725/microsoft-is-doing-the-splits-to-focus-on-ai-and-the-cloud/.

55 Brad Smith and Carol Ann Browne, *Tools and Weapons*, Penguin Random House, 2019, https://www.penguinrandomhouse.com/books/604709/tools-and-weapons-by-brad-smith-and-carol-ann-browne/.

56 Alec Radford et al., “Better Language Models and Their Implications,” OpenAI, February 14, 2019, https://openai.com/blog/better-language-models/.

57 “Code for the paper ‘Language Models are Unsupervised Multitask Learners,’” *GitHub*, 2019, https://github.com/openai/gpt-2.

58 Radford et al., “Language Models are Unsupervised Multitask Learners,” February 2019, https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

59 Rebecca Crootof, “Artificial Intelligence Research Needs Responsible Publication Norms,” *Lawfare*, October 24, 2019, https://www.lawfareblog.com/artificial-intelligence-research-needs-responsible-publication-norms.

60 Toby Shevlane and Allan Dafoe, “The Offense-Defense Balance of Scientific Knowledge: Does Publishing AI Research Reduce Misuse?” In Proceedings of the 2020 AAAI/ACM Conference on AI, Ethics, and Society (AIES ’20), February 7–8, 2020, https://arxiv.org/pdf/2001.00463.pdf.

61 *Ibid.*

62 “OpenAI Charter,” OpenAI, April 9, 2018, https://openai.com/charter/.

[63] Karen Hao, "The messy, secretive reality behind OpenAI's bid to save the world," *MIT Technology Review*, February 17, 2020, https://www.technologyreview.com/s/615181/ai-openai-moonshot-elon-musk-sam-altman-greg-brockman-messy-secretive-reality/.

[64] Alec Radford et al., "Better Language Models and Their Implications," OpenAI, February 14, 2019, https://openai.com/blog/better-language-models/#sample2.

[65] Fabienne Lang, "OpenAI's GPT2 Now Writes Scientific Paper Abstracts," *Interesting Engineering*, October 28, 2019, https://interestingengineering.com/openais-gpt2-now-writes-scientific-paper-abstracts.

[66] "GPT-2 Neural Network Poetry," March 3, 2019, https://www.gwern.net/GPT-2

[67] "Talk to Transformer," https://talktotransformer.com/.

[68] Jack Clark, "GPT-2 model card," *GitHub*, November 5, 2019, https://github.com/openai/gpt-2/blob/master/model_card.md.

[69] Margaret Mitchell, et al., "Model Cards for Model Reporting," *arXiv*, October 5, 2018, https://arxiv.org/abs/1810.03993.

[70] "GPT-2: 1.5B Release," OpenAI blog, November 5, 2019, https://openai.com/blog/gpt-2-1-5b-release/.

[71] Irene Solaiman, "Release Strategies and the Social Impacts of Language Models," OpenAI Report, November 2019, https://d4mucfpksywv.cloudfront.net/papers/GPT_2_Report.pdf.

[72] Jack Clark, "GPT-2 model card," *GitHub*, November 5, 2019, https://github.com/openai/gpt-2/blob/master/model_card.md.

[73] Irene Solaiman et al., "Release Strategies and the Social Impacts of Language Models," OpenAI Report, November 2019, https://arxiv.org/pdf/1908.09203.pdf.

[74] Clément Delangue, "Ethical analysis of the open-sourcing of a state-of-the-art conversational AI," *Medium*, May 9, 2019, https://medium.com/huggingface/ethical-analysis-of-the-open-sourcing-of-a-state-of-the-art-conversational-ai-852113c324b2.

[75] *Ibid*.

[76] Lav R. Varshney et al., "Pretrained AI Models: Performativity, Mobility, and Change," *arXiv*, September 7, 2019, https://arxiv.org/abs/1909.03290.

[77] Rowan Zellers, "Why We Released Grover," *The Gradient*, July 15, 2019, https://thegradient.pub/why-we-released-grover/.

[78] Sajidur Rahman et al., "Attention is All They Need: Combatting Social Media Information Operations With Neural Language Models," FireEye Threat Research, November 14, 2019, https://www.fireeye.com/blog/threat-research/2019/11/combatting-social-media-information-operations-neural-language-models.html.

[79] "The AI On Values," 2017, https://ai-on.org/about/.

[80] Sonnenburg et al., “The Need for Open Source Software in Machine Learning,” *Journal of Machine Learning Research* 8 (2007) 2443-2466, October 2007, http://www.jmlr.org/papers/volume8/sonnenburg07a/sonnenburg07a.pdf.

[81] Karen Hao, “The messy, secretive reality behind OpenAI’s bid to save the world,” *MIT Technology Review*, February 17, 2020, https://www.technologyreview.com/s/615181/ai-openai-moonshot-elon-musk-sam-altman-greg-brockman-messy-secretive-reality/.

[82] Anima Anandkumar, Twitter, February 14, 2019, https://twitter.com/AnimaAnandkumar/status/1096209990916833280.

[83] Vanya Cohen, “OpenGPT-2: We Replicated GPT-2 Because You Can Too,” *Medium*, August 22, 2019, https://blog.usejournal.com/opengpt-2-we-replicated-gpt-2-because-you-can-too-45e34e6d36dc.

[84] Rebecca Crootof, “Artificial Intelligence Research Needs Responsible Publication Norms,” *Lawfare*, October 24, 2019, https://www.lawfareblog.com/artificial-intelligence-research-needs-responsible-publication-norms.

[85] Claire Leibowicz, Steven Adler, and Peter Eckersley, “When Is It Appropriate to Publish High-Stakes AI Research?” The Partnership on AI, April 2, 2019, https://www.partnershiponai.org/when-is-it-appropriate-to-publish-high-stakes-ai-research/.

[86] Paul Lewis, “‘Fiction is outperforming reality’: how YouTube’s algorithm distorts truth,” *The Guardian*, February 2, 2018, https://www.theguardian.com/technology/2018/feb/02/how-youtubes-algorithm-distorts-truth.

[87] Giorgio Patrini, “Mapping the Deepfake Landscape,” Deeptrace, July 10, 2019, https://deeptracelabs.com/mapping-the-deepfake-landscape/.

[88] Xiaoyong Yuan et al., “Adversarial Examples: Attacks and Defenses for Deep Learning,” *arXiv*, July 7, 2018, https://arxiv.org/abs/1712.07107.

[89] Daisuke Wakabayashi, “Self-Driving Uber Car Kills Pedestrian in Arizona, Where Robots Roam,” *The New York Times*, March 19, 2018, https://www.nytimes.com/2018/03/19/technology/uber-driverless-fatality.html.

[90] Irene Solaiman et al., “Release Strategies and the Social Impacts of Language Models,” OpenAI Report, November 2019, https://arxiv.org/pdf/1908.09203.pdf.

[91] Daniel Adiwardana and Thang Luong, “Towards a Conversational Agent that Can Chat About…Anything,” Google AI Blog, January 28, 2020, https://ai.googleblog.com/2020/01/towards-conversational-agent-that-can.html.

[92] Jack Clark, “Import AI: 183: Curve-fitting conversation with Meena; GANs show us our climate change future; and what compute-data arbitrage means,” Import AI, February 3, 2020, https://jack-clark.net/2020/02/03/import-ai-183-curve-fitting-conversation-with-meena-gans-show-us-our-climate-change-future-and-what-compute-data-arbitrage-means/.

[93] Nitish Shirish Keskar et al., “CTRL: A Conditional Transformer Language Model for Controllable Generation,” *arXiv*, September 20, 2019, https://arxiv.org/pdf/1909.05858.pdf.

[94] Lav R. Varshney et al., “Pretrained AI Models: Performativity, Mobility, and Change,” *arXiv*, September 7, 2019, https://arxiv.org/pdf/1909.03290.pdf.

[95] Nate Soares, “2018 Update: Our New Research Directions,” MIRI Strategy, November 22, 2018, https://intelligence.org/2018/11/22/2018-update-our-new-research-directions/#section3.

[96] Kaveh Waddell, “Breaking with tradition, AI research group goes radio silent,” *Axios*, November 27, 2018, https://www.axios.com/artificial-intelligence-research-radio-silent-ed6af5a7-bbb0-46eb-a390-c0fda07cc111.html.

[97] Toby Shevlane and Allan Dafoe, “The Offense-Defense Balance of Scientific Knowledge: Does Publishing AI Research Reduce Misuse?” In Proceedings of the 2020 AAAI/ACM Conference on AI, Ethics, and Society (AIES ’20), February 7–8, 2020, https://arxiv.org/pdf/2001.00463.pdf.

[98] “Secretary Perry Addresses the National Security Commission on Artificial Intelligence,” Department of Energy, November 5, 2019, https://www.energy.gov/articles/secretary-perry-addresses-national-security-commission-artificial-intelligence.

[99] Stephen Cave and Seán ÓhÉigeartaigh, “An AI Race for Strategic Advantage: Rhetoric and Risks,” *AI Ethics and Society 2018*, Volume: 1, 2018, https://www.researchgate.net/publication/330280774_An_AI_Race_for_Strategic_Advantage_Rhetoric_and_Risks/citation/download.

[100] “Forty-two countries adopt new OECD Principles on Artificial Intelligence,” OECD, May 22, 2019, https://www.oecd.org/science/forty-two-countries-adopt-new-oecd-principles-on-artificial-intelligence.htm.

[101] “G20 AI Principles,” G20, 2019, http://k1.caict.ac.cn/yjts/qqzkgz/zksl/201906/P020190610727837364163.pdf.

[102] “What is the OECD?” *The Economist*, July 6, 2017, https://www.economist.com/the-economist-explains/2017/07/05/what-is-the-oecd.

[103] Philip Pierros and Philip Wegmann, “The OECD: New wings or still the same old club?” OECD Observer, http://oecdobserver.org/news/fullstory.php/aid/5698/The_OECD:_New_wings_or_still_the_same_old_club_.html.

[104] Angela Daly et al., “Artificial Intelligence Governance and Ethics: Global Perspectives,” *arXiv*, June 28, 2019, https://arxiv.org/pdf/1907.03848.pdf.

[105] “Scoping the OECD AI Principles,” OECD Digital Economy Papers, No. 291, OECD Publishing, November 2019, https://www.oecd-ilibrary.org/docserver/d62f618a-en.pdf.

[106] “Conference on Artificial Intelligence - AI: Intelligent Machines, Smart Policies,” Going Digital, OECD, October 2017, https://www.oecd.org/going-digital/ai-intelligent-machines-smart-policies/.

[107] Michael Kratsios, “White House OSTP’s Michael Kratsios Keynote on AI Next Steps,” U.S. Mission to the Organization For Economic Cooperation & Development, May 21, 2019, https://usoecd.usmission.gov/white-house-ostps-michael-kratsios-keynote-on-ai-next-steps/.

[108] “G20 AI Principles,” G20, 2019, http://k1.caict.ac.cn/yjts/qqzkgz/zksl/201906/P020190610727837364163.pdf.

[109] Rebecca M. Nelson, “The G-20 and International Economic Cooperation: Background and Implications for Congress,” Congressional Research Service, September 10, 2018, https://fas.org/sgp/crs/row/R40977.pdf.

[110] “G20 Ministerial Statement on Trade and Digital Economy,” June 2019, https://www.mofa.go.jp/files/000486596.pdf.

[111] “OECD AI Policy Observatory,” OECD, Going Digital, September 2019, https://www.oecd.org/going-digital/ai/about-the-oecd-ai-policy-observatory.pdf.

112 Karine Perset, "IGF 2019 - Day 1 - Estrel Saal C - OF39 Artificial Intelligence," Internet Governance Forum, November 26, 2019, https://www.youtube.com/watch?v=6cAqHFJKFD0.

113 "Official Launch of the OECD.AI Policy Observatory: A platform to shape and share AI policies," February 27, 2020, http://oecd.ai/.

114 Richard L. Hudson, "France and Canada move forward with plans for global AI expert council," *Science Business*, November 19, 2019, https://sciencebusiness.net/news/france-and-canada-move-forward-plans-global-ai-expert-council.

115 "OECD Network of Experts on AI (ONE AI)" OECD AI Policy Observatory, OECD, https://oecd.ai/network-of-experts.

116 Tambiama Madiega, "EU guidelines on ethics in artificial intelligence: Context and implementation," European Parliamentary Research Service, September 2019, https://www.europarl.europa.eu/RegData/etudes/BRIE/2019/640163/EPRS_BRI(2019)640163_EN.pdf.

117 "U.S. Leadership in AI: A Plan for Federal Engagement in Developing Technical Standards and Related Tools," US Department of Commerce National Institute of Standards and Technology, August 9, 2019, https://www.nist.gov/system/files/documents/2019/08/10/ai_standards_fedengagement_plan_9aug2019.pdf.

118 "Overview of Saudi Arabia's 2020 G20 Presidency," Saudi Arabia G20, December 2019, https://g20.org/en/g20/Documents/Presidency%20Agenda.pdf.

CLTC

Center for Long-Term Cybersecurity

UC Berkeley

Center for Long-Term Cybersecurity
cltc.berkeley.edu
@CLTCBerkeley