Evaluation of the Ethics and Governance of Artificial Intelligence Initiative

Caribou Digital | May 2022

The Miami Foundation
For A Greater Miami

Luminate
Founded by THE OMIDYAR GROUP

Reid Hoffman

KNIGHT FOUNDATION

WILLIAM + FLORA Hewlett Foundation

Acknowledgements

This evaluation report was written by Troy Etulain, Niamh Barry,* and Chelsea Horváth, with support from Jonathan Donner. Special thanks to Shelby Brewster for editing, Tom McEvoy for design, and Lindsey Linzer at The Miami Foundation and John Sands at Knight Foundation for their constant availability, assistance, and enthusiasm throughout the process. This evaluation depended on the generosity of all individuals who took time to speak with us about the Initiative; thank you.

The views presented in this report are those of the authors and do not necessarily represent the views of The Miami Foundation.

*For questions, please contact niamh@cariboudigital.net.

Caribou Digital is a research and advisory firm that seeks to change the world by helping organizations build inclusive and ethical digital economies. All Caribou Digital reports are available at www.cariboudigital.net

Table of contents

Acronyms

ACLU	American Civil Liberties Union
ACM FAccT	Association for Computing Machinery Conference on Fairness, Accountability, and Transparency
AI	Artificial intelligence
AIA	Algorithmic Impact Assessments
BayesDB	Bayesian database
BBC	British Broadcasting Corporation
BKC	Harvard Berkman Klein Center for Internet & Society
COGSEC	Cognitive Security Conference
COVID-19	Coronavirus
CUNY	The City University of New York
EFF	Electronic Frontier Foundation
EGAI	Ethics and Governance of Artificial Intelligence Initiative
EU	European Union
EUR	Europe
FAT ML	Fairness, Accountability, and Transparency in Machine Learning Conference
GDPR	General Data Protection Regulation
GovAI	Centre for the Governance of AI
HRDAG	Human Rights Data Analysis Group
HSBC	Hong Kong and Shanghai Banking Corporation
ICRC	International Committee of the Red Cross
IEEE	Institute of Electrical and Electronics Engineers
ITS	Institute for Technology and Society of Rio de Janeiro
MIT	Massachusetts Institute of Technology
ML	Machine learning
NDA	Non-disclosure agreement
NYPD	New York City Police Department
NYU	New York University
OECD	Organisation for Economic Co-operation and Development
OII	Oxford Internet Institute
PAI	Partnership on AI
PILAC	Harvard Law School Program on International Law and Armed Conflict
POST	Public Oversight of Surveillance Technology
RFP	Request for proposals
RIT	Rochester Institute of Technology
SEAS	Harvard John A. Paulson School of Engineering and Applied Sciences
SIPRI	Stockholm International Peace Research Institute
UK	United Kingdom
UN	United Nations
UNICEF	United Nations Children’s Fund
US	United States
UT	University of Texas

Executive summary

In 2016, as the Ethics and Governance of AI Initiative (the Initiative) was being conceptualized, numerous events occurred that would impact research, policy, and public discourse on the ethics and governance of AI. Examples include: the founding of the Partnership on AI (PAI); the ProPublica investigation that uncovered significant racial bias in AI used by law enforcement; and Brexit and the US presidential election, two political events which involved the spreading of misinformation on social media platforms.[1] The funders recalled that the field of AI ethics was nascent when the Initiative was created: "There was definitely a sense in 2016 that there was so much going on [...] it was a very rapidly moving field that hadn't taken shape at all."[2]

By 2017, $26 million had been raised for the Initiative that sought "to ensure that technologies of automation and machine learning are researched, developed, and deployed in a way which vindicates social values of fairness, human autonomy, and justice."[3] Philanthropic support was provided by Luminate (founded by The Omidyar Group), Reid Hoffman, Knight Foundation, and the William and Flora Hewlett Foundation. The Miami Foundation provided fiscal management. The Initiative was structured as a joint project of the MIT Media Lab and the Harvard Berkman-Klein Center for Internet and Society (BKC).

In their joint proposal, BKC and Media Lab articulated that, in collaboration with partners, they would "deploy new prototypes, conduct research, directly impact both policy and technologies, build community, teams, and even institutions, and engage in education and outreach that meaningfully connects human values with the technical capabilities of AI...."[4] The Initiative was active from 2017 to 2022 and awarded approximately $23 million to 39 grantees working on 42 projects.

As the Initiative neared the end of its funding, The Miami Foundation and funding partners sought to assess the durability of its collaborative efforts, and the impact of projects supported through its grants. In August 2021, The Miami Foundation contracted Caribou Digital to evaluate the Initiative by reviewing 200+ Initiative documents, surveying grantees, and conducting 30 interviews with Initiative stakeholders. The top-level findings and recommendations from this evaluation are presented in this report.

Initiative impact

Using the Initiative's implied Theory of Change—reconstructed by Caribou Digital in Annex 3—as a framework, impact was described and assessed across four categories: 1) relevance and centrality of assets developed under the Initiative, 2) informed public and private sectors, 3) changes in governance, public policy, and industry practice, and 4) building the AI ethics and governance community.

The Initiative generated vast quantities of assets: over 250 publications (Annex 6), more than a dozen products (Annex 5), and countless engagements. In terms of their centrality and utility to the broader field, academic citations of these assets ranged from zero to thousands. Uptake of AI products for public good varied, with some stand-out examples of high and sustained uptake. Insights from large events funded by the Initiative suggested high relevance, engagement, and value.

One in three of the Initiative's grantees provided examples of their contributions to more informed public and private sectors. Grantees informed policymakers in a variety of ways: providing

[1] Sam Levin and Nicky Woolf, "Tesla Driver Killed while Using Autopilot was Watching Harry Potter, Witness Says," *The Guardian*, July 1, 2016, https://www.theguardian.com/technology/2016/jul/01/tesla-driver-killed-autopilot-self-driving-car-harry-potter; Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner, "Machine Bias," ProPublica, May 23, 2016, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.
[2] Martin Tisné (Luminate), online interview, January 12, 2022.
[3] "Request for Proposals: Assessing the Ethics and Governance of Artificial Intelligence Initiative."
[4] BKC-Media Lab, Cover Memo to Ethics and Governance of AI Fund Principals, June 1, 2017.

evidence, testifying, delivering briefings, sitting on advisory groups, and engaging in partnerships. Industry representatives were notably more difficult to engage, as they were less accessible and less likely to share that a grantee's work was informative. Excluding a number of internationally mandated institutions, the majority of examples emanated from North America, Europe, and the UK.

One in four of the Initiative's grantees linked explicit policy changes and actions to their work. Grantees identified changes at several major technology companies, including: improvements to the quality of information on platforms (Twitter, Pinterest, Google, Facebook); online safety protections for users (Disqus); and assessments on bias within AI systems (Amazon and HSBC). However, grantees highlighted that impact on technology companies may be underreported. Within the public sector, all changes were concentrated in the US, at the local level, such as the ban on face surveillance technology in Massachusetts, and Electronic Frontier Foundation's (EFF) work on the Public Oversight of Surveillance Technology (POST) Act in New York City. EFF's two legal rulings to reverse the use of AI to implicate or imprison, set precedents for future campaigns.

Views about the cohesiveness and strength of the AI ethics and governance community varied considerably, as did assessments of the roles of BKC, MIT, and the Initiative in strengthening them. Some noted that it "definitely exists" and that BKC and MIT "definitely contributed to it." Others thought that, while there is a "healthy field," the unification of computer science and social science "hasn't ended up with that galvanization." But one thing is certain; the number of institutions producing outputs and the number of people convened under the banner of AI ethics grew, and the Initiative fueled this growth.

Two broader changes enabled by the Initiative surfaced: organization progression and career progression and change. Some research products produced through the Initiative contributed to their authors' career progression from research to developing policy or practices around AI ethics and governance. A few grantees credited the Initiative with their growth from projects to organizations and as leaders in their field. For example, the Markup used Initiative funds as seed funding; by the end of their grant, they had raised $25 million.[5] DigiChina transitioned from a startup project within the New America Foundation to a program based at Stanford University with multi-year funding[6]

Unsurprisingly for an initiative of this scale and diversity, many projects continued to generate impact and a few closed. The benefits of educated professionals, members of the public, research assets, and legal reforms are relatively durable. Some Initiative projects, such as Tattle, the Markup, DigiChina, CivilServant and the FAT ML (later the ACM FAccT) conference, have grown or found homes in new institutions and continue to add value. However, 18% (n=7) of grantees, representing 6% ($1,367,188) of total funding, did not report on impact beyond outputs.

Was aggregate impact observed enough? While it may be up to each funder to assess whether reported impact was sufficient, it is also worth considering the nascency of AI ethics and governance in 2016, which required an element of foundation laying. This more exploratory work tends to weigh towards outputs rather than longer-term impacts. Ultimately, Initiative leadership and funders should emerge with new knowledge and a clearer view on where resources should be focused next or new learnings to apply to similar future initiatives.

Initiative implementation

There was a relatively proportional mix of theoretical and practical projects. There were significantly more research papers produced than AI products for public good developed. However, if a broader view of "practical" is taken—i.e., including engagement with public and private sector actors and the development of public resources and trainings—the theoretical and practical mix does not appear disproportional. Further, 72% of grantees worked in more than one of the Initiative's three strategic areas—1) community and capacity building, 2) research sprints and pilot projects, and 3)education, training, and outreach—demonstrating that most projects embodied a mix of theoretical and practical.

[5] The Markup, final report, October 2019.

[6] New America, final report, October 2021.

The Initiative was responsive to most trends in the broad field of AI. This responsiveness can be characterized across four efforts: 1) to embrace the inherent interdisciplinarity of AI (and of AI ethics and governance); 2) to uplift and amplify a diversity of voices; 3) to include and engage broader elements of society; and 4) to develop and support a counterweight to industry resources and priorities

> **Interdisciplinarity**. It is notable, responsive, and appropriate that the majority of the Initiative projects had various interdisciplinary aspects to them—either in their teams or in the people they convened. Grantees felt that such interdisciplinarity was important to continue and in the long term the community will be healthier and more resilient for this.
>
> **Diversity.** While the "diversity disaster" in the broader AI field is well known.[7] Within the sub-field of AI ethics and governance, grantees noted that a field that relied on the same voices, geographies and, often, institutions would result in missing perspectives. With 15% of grantees being non-US based[8] and 80% of funding support academic institutions, on this front, and in line with their international ambitions[9] the Initiative could have done more. There remains an imperative to push for substantive diversity of thought and experience, both within geographies and across them.
>
> **Active inclusion of society.** Over 50% of the Initiative's projects included society as one of their target audiences. Several grantees shared that the AI ethics and governance community must build on the Initiative's efforts to actively engage with society. Democratic societies determine how their governments use technology and automated systems; the Initiative illustrates how it is vital to bridge the knowledge gap and explain how these systems work to ensure that current social injustices are not replicated through AI.
>
> **Industry counterweight.** While the Initiative may be seen as a counter-weight to the significant industry resources invested in AI. Donors have the opportunity to off-set market and geopolitical incentives in support of human-centric and ethical applications of AI. While it will not be possible for donors' funds to equal the amount spent on AI by industry or major governments, academic and civil society organizations will continue to play a key role in increasing public awareness and influencing policy.

Recommendations for funders

The three most pertinent recommendations to support broad, complex multi-donor/ year/grantee initiatives are highlighted here.

1. **Design for diversity—in institutions, approaches, and geography—in the selection process.** Conducting informative activities—such as ecosystem scanning and surveys on priorities—prior to selection processes is an opportunity to gain consensus on the gaps in research and practice and increase awareness of a broader range of organizations conducting relevant work. Another approach is to set a quota/cap on the number of grants provided to certain organizational types, those from specific countries, or those representing specific interests. This will enable a richer set of implementers, perspectives, and impacts. Support for intra-initiative engagement—from internal newsletters, discussion forums, or annual convenings—could further maximize the benefits of diverse grantee voices.
2. **Define the community mission to galvanize people and institutions.** Building communities is inherently difficult work, made more difficult without clarity on the community's mission. For future community-building initiatives, articulating the vision and mission, clarifying membership, and determining strategies to achieve the mission would galvanize people and institutions towards it.
3. **Design for impact measurement at the start of initiatives.** A framework could include: a robust theory of change, measurement principles, specific and appropriate metrics, dedicated resources to regularly aggregate and review insights generated by grantees, and intra-Initiative learning convenings.

[7] West, S.M., Whittaker, M. and Crawford, K. Discriminating Systems: Gender, Race and Power in AI. AI Now Institute. April 2019. Retrieved from https://ainowinstitute.org/ discriminatingsystems.html

[8] South America (2), UK (2), and Asia (2)

[9] BKC and Media Lab, proposal narrative, April 12, 2017.

Introduction

In 2016, as the Initiative was being conceptualized, numerous events occurred that would impact research, policy, and public discourse on the ethics and governance of AI. These included the founding of the **Partnership on AI** (PAI), the first fatal crash of a car driving on autopilot, the ProPublica investigation that uncovered significant racial bias in algorithms used by law enforcement, and Brexit and the US presidential election, two political events which involved the spreading of misinformation on social media platforms.[10] The funders recalled that the field of AI ethics was nascent when the Initiative was created: "There was definitely a sense in 2016 that there was so much going on [...] it was a very rapidly moving field that hadn't taken shape at all."[11]

By 2017, $26 million had been raised for the Initiative. This included $10 million each from **Luminate** (founded by **The Omidyar Group)** and **Reid Hoffman**, $5 million from **Knight Foundation**, and $1 million from the **William and Flora Hewlett Foundation**. The Miami Foundation provided fiscal management.

The Initiative was conceived as a hybrid research effort and philanthropic fund that sought "to ensure that technologies of automation and machine learning are researched, developed, and deployed in a way which vindicates social values of fairness, human autonomy, and justice."[12] As a joint project of the **MIT Media Lab** and the **Harvard Berkman-Klein Center for Internet and Society (BKC)**, the Initiative intended to build pathways for collaboration across **Media Lab** and **BKC** and incubate a range of research, prototyping, and advocacy activities within these two anchor institutions and across the broader ecosystem. The Initiative was active from 2017 to 2022 and awarded approximately $23 million to 39 institutions working on 42 projects.

Evaluation methods

The Miami Foundation and the funding partners sought to assess and understand the impact of the Initiative. In August 2021, The Miami Foundation contracted Caribou Digital to evaluate the Initiative.

The evaluation was conducted in three phases: 1) inception, 2) data collection, and 3) analysis. During the inception phase, the evaluation team worked to fully understand the original vision of the Initiative, including its implicit Theory of Change. The data collection phase focused on specific questions related to the structure, implementation and impact of the Initiative. The evaluation was a qualitative design using document review, semi-structured interviews, and surveys. See Annex 1 for a complete description of the evaluation framework and data collection and analysis methods.

Report outline

1. **Initiative implementation**. Outlines the governance structure, processes and outcomes of project selection. It also reflects on the Initiative's responsiveness to trends in the field of AI.
2. **Initiative impact**. Describes and assesses four categories of impact that the Initiative sought to effect.
3. **Recommendations**. Outlines three broad recommendations for funders supporting or interested in launching similar initiatives.
4. **In summary.** Summary reflections on the aggregate achievements of the Initiative.

[10] Sam Levin and Nicky Woolf, "Tesla Driver Killed while Using Autopilot was Watching Harry Potter, Witness Says," *The Guardian*, July 1, 2016, https://www.theguardian.com/technology/2016/jul/01/tesla-driver-killed-autopilot-self-driving-car-harry-potter; Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner, "Machine Bias," ProPublica, May 23, 2016, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.
[11] Martin Tisné (Luminate), online interview, January 12, 2022.
[12] "Request for Proposals: Assessing the Ethics and Governance of Artificial Intelligence Initiative."

Initiative implementation

This section outlines the implementation of the Initiative, specifically the 1) project selection processes, 2) type of projects funded, and 3) a review of the Initiative's responsiveness to trends in the AI field.

Project selection process

A Steering Committee was established in October 2017 to advise on processes and grantee selection. It consisted of 18 members of which 6 were voting members. The Steering Committee hired an Initiative director, Tim Hwang, to coordinate project selection with **BKC** and **Media Lab** and oversee all grants.

The project selection process was intended to reflect a "Silicon Valley mindset" i.e., making decisions quickly and in response to opportunities or challenges observed, without an overly formal selection process. There were three modes in which awards were provided: 1) annual grants to **BKC** and **Media Lab**, 2) network grants, and 3) a Challenge Fund. Please see Annex 2 for a breakdown of the Initiative granting timeline and budget allocation.

1. **Annual grants to anchor institutions:** The Initiative provided annual grants to **BKC** and **Media Lab** to fund various existing and new in-house projects. **BKC** received funding for three years (2017–2019) and Media Lab for two years (2017–2018). In total, five awards were provided to **BKC** and **Media Lab** valued at $14,482,436, with an average grant size of ~$2.8 million.
2. **Network grants:** Recipients of these grants were sourced via the Initiative director and the anchor institutions' networks in conformance with a desire to quickly fund known opportunities. These totaled 29 awards valued at $8,093,448, with an average grant size of ~$250,000.
3. **Challenge Fund grants:** The AI and the News Challenge was an open call in 2019. According to Initiative Director Tim Hwang, the challenge grants tended to be "smaller, more experimental, not necessarily associated with established organizations. They were designed to be awarded quickly and intended to be 'prototype-style projects.'"[13] These totaled seven awards valued at $750,000, with an average grant size of ~$100,000.

Typologies of projects: Theoretical and practical

In this section, the types of projects chosen via the various selection processes are reviewed.

In their joint proposal, **BKC** and **Media Lab** outlined three broad strategic activities: 1) Community and capacity building, 2) Research sprints and pilot projects and 3) Education, training, and outreach. When all projects are plotted against these activities, insights into the Initiative's aggregate focus are revealed (Figure 1). Research and pilot projects were a central pillar, with 87% (n=34) of grantees recording related activities. This was followed by 64% (n=25) of grantees supporting education, training, and outreach activities and 56% supporting community building (n=25). The majority of grantees, 72% (n=28) worked in at least two activities, with 28% (n=11) dedicating resources to a single activity area.

A sample of a significant project grouping is described below, and a description of all projects can be found in Annex 4.

[13] Tim Hwang (previous director of the Initiative), online interview, November 11, 2021.

Figure 1: Plotting projects against the expanded Initative's strategic pillars

**Ordering of funded projects/institutions is chronological based on funding start date*

		Community and capacity building			Research sprints and pilot projects				Education, training, and outreach						
Organization or project name	Funding received	Events | working groups | community spaces	Fellowships	Supporting establishment or continued work of mission aligned institutions	Product/ Infrastructure development	Applied research/ experiments	Research	Journalism	Policy engagement (public bodies)	Industry engagement	Civil society engagement	Public campaigns/ awareness raising	University courses	Training programs	Practical resources for outreach/ training
Harvard; Year 1, 2 & 3	$8,314,773	●	●	●		●	●		●	●		●	●	●	●
MIT; Year 1 & 2	$6,167,663	●			●	●	●		●	●		●	●	●	
Access Now	$200,000	●					●		●			●			●
New York University; AI Now	$662,000	●					●		●						
Data & Society Research Institute	$200,000	●		●			●		●	●					
The Institute for Technology & Society of Rio	$280,021	●			●		●						●	●	●
University of Utah	$59,897				●		●								●
Cornell University; FAT ML	$166,714	●													
Cambridge in America; Leverhulme Centre	$252,291	●				●	●								
Digital Asia Hub	$100,000	●	●	●			●				●		●		
Mozilla Foundation	$250,000		●				●		●	●					
Meedan	$175,000	●			●	●			●	●					
New America Foundation	$250,000		●				●		●						
The Markup	$750,000			●		●		●				●			
Community Partners (HRDAG)	$300,000					●									
ACLU Foundation of Massachusetts	$500,000						●		●		●	●			●
Princeton CITP	$201,840	●					●		●	●			●		
University of Regina	$275,000					●	●			●		●			
Harvard; SEAS	$135,000					●									
Harvard; Assembly Accelerator	$60,000	●	●	●	●	●									
Harvard; PILAC	$280,685	●					●		●		●				
University of California, Berkeley	$400,000					●									
Harvard; EJ Safra Center	$120,000		●				●								
GovAI	$250,000						●								
Global Voices; Civil Servant	$275,000				●	●			●	●	●	●			
Harvard; Shorenstein Center	$700,000		●				●		●	●	●	●		●	●
Tattle Civic Technologies	$100,000				●		●								●
Legal Robot	$100,000				●										
CUNY	$100,000		●											●	●
MuckRock	$150,000				●										
Chequeado	$75,000						●	●				●			●
Seattle Times	$125,000							●				●		●	●
RIT	$100,000				●						●				●
Electronic Frontier Foundation	$235,000						●		●			●			
Data Nutrition Project	$230,000				●		●								
Oxford Internet Institute	$365,000					●	●		●	●		●			
UT-Austin	$200,000	●													
Tim Hwang	$200,000	●													
WeRobot	$20,000	●													
Total	$23,325,884	16	8	5	11	12	23	3	15	10	6	12	5	6	11

Academic institutions

Academic institutions represented the majority of grantees (46%; n=18) and received 80% of the total funding. Significant volumes of research were generated by academic institutions and topics spanned misinformation, governance, privacy, accountability, human rights, criminal justice, health, and armed conflict. Principles in AI development were aggregated and new methodologies to assess AI impact were forwarded. Some institutions developed AI products; for example, **ITS Rio** developed bots to locate other bots on social media, and **RIT** developed a product to assess deepfakes. In aggregate, these institutions convened numerous workshops, developed community spaces, hosted large cross-disciplinary events, and supported fellowships. Collectively, the public sector (i.e., legal, health, military, governance), civil society, international bodies, and industry were engaged to discuss the ethics and governance of AI. Training opportunities and courses were developed and deployed. Within academic institutions a particular genre of academics were noted to be highly active in promoting their research beyond journals and conferences and sought wider media coverage. Such promotion efforts supported industry and policy engagement. Their research methods ranged from innovative citizen science (**CivilServant**) to behavioral science (**University of Regina**).

Research organizations and think tanks

The five research/think tank organizations—**Mozilla, Data & Society, New America Foundation, Cambridge in America Leverhulme Centre,** and **Digital Asia Hub**—cut across all three activity areas, but conducting research and using the outputs for policy and industry outreach were their primary activities. Research topics were diverse and included understanding the role of AI in disinformation, the social and economic implications of AI, and geo-targeted research on AI in Asia. All but one organization engaged with government policymakers. Two organizations engaged with technology companies, while another specifically engaged in public outreach. These five organizations received 5% of the total funding.

Technology companies

Four technology companies were funded. Most had a strong information quality/media focus. **Meedan** developed products to assess article credibility alongside AI fact-checking newsroom products. **Tattle Technologies** also supported fact-checking, and two grantees (**Legal Robot** and **Muckrock**) developed AI products to acquire and classify large public datasets to enable greater access to information for civil society and the media. These four organizations received 2% of the total funding.

Advocacy

Four advocacy organizations pursued research to inform and further their advocacy efforts on: transparency and explanation of machine learning (ML) (**EFF**); misuse of face surveillance technology (**ACLU**); Europe's GDPR and e-privacy reforms (**Access Now**); and AI-generated pretrial risk assessments (**Human Rights Data Analysis Group**). They produced materials for the general public and legal sector, organized public awareness campaigns, and directly engaged with policymakers. These three organizations received 5% of the total funding.

Media

Three media outlets were supported to seed or produce investigative journalism on topics related to ethics and governance of AI, which focused on audiences in the US (**Seattle Times** and **The Markup** primarily) and Latin America (**Chequeado**). Public outreach was via their own channels and other media outlets. These three organizations received 4% of the total funding.

Note that there were five projects supported that did not fit the classification above; these were mainly standalone research projects and events.

While academic institutions were the majority of funding recipients, research was not the exclusive output of their projects. Outputs ranged from product development, convenings, training resources, and engagement with the public sector, private sector, and general public. Thus the Initiative embodied a mix of practical and theoretical projects. Ultimately, the grantees used varying *levers*, mostly based on product and knowledge assets produced, to support the ethical advancement of AI. Some interviewees noted that the diversity was intentional and likened it to a series of bets or experiments to understand what *might* be the optimum levers that address the ethical concerns of AI. These levers included:

- Supporting journalistic and legal organizations to hold algorithms to account
- Engaging in multidisciplinary convenings and collaborations of ethicists and computer scientists
- Supporting products and methodologies to determine AI fairness, accountability, and transparency
- Education and training opportunities for public institutions and the media
- Investment in causal reasoning and behavior science to understand engagement with AI
- Public awareness campaigns on specific uses of, or policies around, AI.

Initiative's responsiveness to trends

Through the projects implemented by the Initiative, this section identifies how the Initiative responded to trends in the broad field of AI. This responsiveness can be characterized across four efforts: 1) interdisciplinarity, 2) diversity, 3) active inclusion of society, and 4) supporting a counterweight to industry resources.

The interdisciplinarity of the endeavor: During the period of the Initiative, the field of AI ethics seems to have reached the tipping point of complexity, where experts have lost the ability to keep track of developments.[14] This is more than a challenge of scale—it is one of heterogeneity of approaches. Several grantees mentioned the continued need for the AI ethics and governance community to be interdisciplinary. It is notable, responsive, and appropriate that the majority of the Initiative projects had various interdisciplinary aspects to them—either in their teams or in the people they convened. Grantees felt that such interdisciplinarity was important to continue. Dr. Sandra Wachter from **OII** shared:

> "I think it's really important that the focus is not just on one particular discipline at the moment. [...] There is no tech problem that doesn't have a social problem underneath it, right? [...] So funding definitely needs to be put into a direction that encourages working together."[15]

Strength in diversity: In 2019, **AI Now** published a study that highlighted the "diversity disaster" in the field of AI, and flagged that the biases of systems built by the AI industry can be largely attributed to the lack of diversity within the field itself.[16] Within the sub-field of AI ethics and governance, grantees noted that a field that relied on the same voices, geographies and, often, institutions would result in missing perspectives. **ITS Rio** stated that:

> "The policy debates about AI have been predominantly dominated by organizations and actors in the Global North. [...]we have noticed that there is a growing need for a more diverse perspective regarding the policy issues and consequences of AI, especially in the Global South."[17]

In addition, Eric Sears from the **MacArthur Foundation** shared:

> "In the AI field there are many fellowship opportunities geared towards people who are already in the field, already succeeding financially, and are privileged in other ways. There is a lack of fellowship opportunities that center the needs of people from historically marginalized communities to help ensure they have what they need to succeed [...]There is a big opportunity for philanthropy to help shift this and advance the AI field in a more equitable direction."[18]

With 15% of grantees being non-US based and 80% of funding support academic institutions, on this front, and in line with their international ambitions[19] the Initiative could have done more. There

[14] Cornell University FAT ML/ACM FAccT, Caribou Digital-administered online survey, December 6, 2021.
[15] Sandra Wachter (Oxford Internet Institute), online interview, January 25, 2022.
[16] West, S.M., Whittaker, M. and Crawford, K. Discriminating Systems: Gender, Race and Power in AI. AI Now Institute. April 2019. Retrieved from https://ainowinstitute.org/ discriminatingsystems.html
[17] ITS Rio, final report, July 2018.
[18] Eric Sears (MacArthur Foundation), online interview, February 3, 2022.
[19] BKC and Media Lab, proposal narrative, April 12, 2017.

remains an imperative to push for substantive diversity of thought and experience, both within geographies and across them.

Active engagement with society: Several grantees, including **EFF**, **ACLU**, **AI Now**, and A**ccess Now**, shared that the AI ethics and governance community must build on the Initiative's efforts to actively engage and reach out to society. Just over half of grantees had the public/civil society as at least one of their target audiences. For some research-focused initiatives, engaging society may require resources, different and complementary capabilities, and more expansive and detailed theories of change, yet it's important for the public to understand the risks and benefits of AI for individuals, communities, and society at large.

Counterweight to industry resources: A 2019 *New York Times* article observed that AI research was becoming increasingly expensive, with the danger that pioneering AI research will be a field of haves and have-nots. The haves will be mainly a few big companies like Google, Amazon, and Facebook. The article quoted concerns from computer science academics that: "the huge computing resources these companies have pose a threat—the universities cannot compete."[20]

PAI was launched in 2016 by a consortium of big technology companies and was chartered to "conduct research, recommend best practices, and publish under an open license in areas such as ethics, fairness and inclusivity; transparency, privacy, and interoperability..."[21] It aimed to have equal representation from corporate and non-corporate organizations. **Access Now** joined **PAI**, but withdrew in 2020, stating in an open letter that it "did not find that **PAI** influenced or changed the attitude of member companies or encouraged them to respond to or consult with civil society on a systematic basis."[22]

Similar themes were forwarded through interviews with grantees and the Initiative leadership; the concept of a counterweight was built into the vision of the Initiative. Jonathan Zittrain, co-founder of **BKC**, raised a similar concern on the direction of AI research, specifically where the greatest advances in AI research will come from: "whoever has the data or the most PhDs is where the advances are going to come from," noting that the number of PHDs supported by big tech companies, outweighs those of academia on similar topics. [23]

To the extent that the $23-million-dollar Initiative can be framed as a donor-led counterweight to the industry's deep resources more generally, the arguments for its creation may still be valid today, particularly as the field has become larger and more complex.

Reflections on Initiative implementation

Theoretical and practical projects supported: There were *significantly* more research papers produced than products developed. Yet, when taking a broader view and including public and private sector engagement and public resources and trainings developed, the theoretical and practical mix does not appear disproportional. As noted, 72% of grantees worked in more than one activity area.

Responsiveness to trends: The Initiative was responsive to most trends in the broad field of AI. 1) The Initiative responded to the need for interdisciplinary teams to address issues of AI ethics and governance by supporting numerous projects and methodologies that encapsulated this trend, 2) strong efforts to actively engage with and reach out to society, 3) the Initiative was a counterweight to industry resources and priorities on AI and, 4) reflecting on geographical and institutional diversity, more could have been done.

[20] Steve Lohr, "At Tech's Leading Edge, Worry About a Concentration of Power," *New York Times*, September 26, 2019, https://www.nytimes.com/2019/09/26/technology/ai-computer-expense.html.

[21] Alex Hern, "'Partnership on AI' Formed by Google, Facebook, Amazon, IBM and Microsoft," *The Guardian*, September 28, 2016, https://www.theguardian.com/technology/2016/sep/28/google-facebook-amazon-ibm-microsoft-partnership-on-ai-tech-firms.

[22] Access Now, "Access Now Resigns from the Partnership on AI," *Access Now* (blog), October 13, 2020, https://www.accessnow.org/access-now-resignation-partnership-on-ai/.

[23] Jonathan Zittrain (BKC), online interview, October 19, 2021.

Initiative impact

With an understanding of the projects implemented under the Initiative and using the Initiative's implied Theory of Change—reconstructed by Caribou Digital in Annex 3—as its framing, this section describes and assesses four categories of impact that the Initiative sought to effect:

1. Relevance and centrality of assets developed under the Initiative
2. An informed public and private sector
3. Changes in governance, public policy and industry practice
4. Building the ethics and governance of AI community

Relevance and centrality of assets developed under the Initiative

The grantees generated vast quantities of diverse assets: over 250 publications, more than a dozen products/services, a number of international convenings and countless meetings, working groups, and one-to-ones with peers, policymakers, tech industry, and civil society. (See Annex 6 for a sample of research outputs and Annex 5 for products developed.) Critical to the Theory of Change is that such outputs are seen as relevant and central enough by their respective audiences to generate uptake and thus add value. Within this impact section, five modes of uptake are described with a *sample* of relevant grantee projects:

1. Research replication and citations
2. Media coverage
3. Adoption of products and services
4. Engagement with events
5. Educational and training programs

Research replication and citations

With the aim to sample the topics explored—rather than serve as a complete bibliography of the Initiative—the evaluation team noted 250+ knowledge assets generated by grantees. Citations serve as an academic metric of understanding centrality. Gordon Pennycook, assistant professor at the **University of Regina,** noted that "the research supported by this grant, over a dozen papers, has been cited well over 2,000 times."[24] Gordon Pennycook also noted that their study on fake news was "one of the first ever studies" to address the topic, and that their methodology has been used in "dozens and dozens" of papers in "lots of different institutions" since publication[25] He reflected that the centrality of their work was due to timing: "You take on a topic that is of clear and present concern that a lot of people are going to be drawing their attention to, and then you give them the tools to be able to investigate it. That's how, and then that's what creates a big impact."[26] A collaboration between **BKC** and **MIT's Center for Civic Media** provides another example. A team of researchers mapped the online media ecosystem during key moments; their research culminated into a book, *Network Propaganda*.[27] **BKC**'s report noted that "research collaborators in locations such as France, Germany, Spain, and Colombia are currently replicating the research methodology outlined in *Network Propaganda* in an effort to better understand the online media ecosystem's impacts during their own elections." [28]

With such quantities of research, the range of citations varied. The evaluation team randomly sampled 45 academic research pieces, and citations ranged from 0 to 985, with 31% being cited five or fewer times and 33% being cited 100 or more times.

[24] University of Regina, Caribou Digital-administered online survey, November 5, 2021.
[25] Gordon Pennycook (University of Regina), online interview, November 17, 2021.
[26] Gordon Pennycook (University of Regina), online interview, November 17, 2021.
[27] Yochai Benkler, Robert Faris, and Hal Roberts, *Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics* (New York: Oxford University Press, 2018).
[28] BKC, year two final report, September 2019.

Media coverage

Media coverage expands the research audience, and the media provides a service in translating complex topics into language accessible to the general public. A number of grantees made a point of highlighting media coverage they received.

- **NYU's AI Now** tracked and reported 63 incidences of media coverage, **Media Lab's Moral Machine** reported 115, and **OII** almost 50 incidences.[29] Numerous top-tier international and national media outlets covered grantees' research, including Forbes, Wired, BBC, Quartz, *Washington Post*, Fortune, Bloomberg, *The Telegraph*, *The Economist*, Nature News, *The New Yorker*, Vice, *Business Insider*, *The Guardian*, *Scientific American*, The Verge, *Der Spiegel* (Germany), *Le Monde* (France), and the Australian Broadcasting Corporation.[30]

- Other notable coverage included: **The Institute for Technology & Society (ITS) Rio** and **Chequeado**'s coverage in Latin America; **ACLU**'s mainstream coverage through their week of action with youth; and **CivilServant**'s feature in a cover story in *The Atlantic*, "How to Put out Democracy's Dumpster Fire."[31]

- Funded media outlets, such as the ***Seattle Times*,** also cited uptake by *other* media outlets and feedback on the value of their content among their readers. In their report they noted that a story about anti-surveillance measures went viral, influencing 11 new subscriptions and landing on the Pocket Hits list as one of the most popular stories in the nation for a week. The story also led to an interview on the Tech News Weekly TV show.[32] Feedback from readers highlighted that new insights were also gained:

> "I'm so thankful our Bloomington paper carried your article about the introduction of robots at Walmart. I had heard vaguely about how robots were coming to a variety of work settings, but from you I was able to learn some specifics about the pros and cons. I appreciate so much how you were able to pull together many viewpoints and perspectives, and could give me a balanced understanding of where we are going."[33]

Adoption of products and services

The evaluation team found references to 21 AI products for public good that were either prototyped or improved under the Initiative—Annex 5. Many of the products developed could be categorized as supporting information quality via enabling access to vast quantities of information, fact-checking, assessing disinformation and fake/bot accounts, and quality-assuring datasets used by AI. A handful of other products and services were created for a research use case, for example, **Bartleby**.[34] Product development/enhancement within existing technology organizations showed examples of high uptake. For example, **MuckRock**'s toolkit of ML classifiers for newsrooms and researchers, has enjoyed large scale uptake:

> "I think if you go into almost any newsroom in the country, people use at least one of our tools. We have 4,000 newsrooms on our platform. We have 70,000 registered users. One of our tools doesn't have a registration component and that runs about 300,000 people a month who

[29] For OII the 50 incidences were reported in 2020 in their interim report, the project ongoing until 2022 and thus this is unlikely to be the total number,

[30] New York University AI Now, final report, March 2019; Oxford Internet Institute, interim report, October 2020; Media Lab, interim report, February 2019.

[31] ACLU Massachusetts, Caribou-Digital administered online survey, November 1, 2021; Chequeado, final report, October 2020; ITS Rio, final report. July 2018; J. Nathan Matias (CivilServant), online interview, December 20, 2021; "Face Surveillance: Student Week of Action," ACLU Massachusetts, https://www.aclum.org/en/face-surveillance-student-week-action; Anne Applebaum and Peter Pomerantsev, "How to Put Out Democracy's Dumpster Fire," The Atlantic, March 8, 2021, https://www.theatlantic.com/magazine/archive/2021/04/the-internet-doesnt-have-to-be-awful/618079/.

[32] *Seattle Times*, final report, June 2020.

[33] *Seattle Times*, final report, June 2020.

[34] Bartleby is software for large-scale management of GDPR, CCPA, and university research ethics procedures.

> actively use that. And so we have just, you know, a very, very broad, broad reach across our portfolio."[35]

Similarly, **Meedan**'s Bot Garden (part of Check, which supports fact-checking at scale) referenced partnerships with numerous media outlets and uptake amongst social media and chat apps.[36] Lastly, **Tattle**'s datasets and tools to enable storytelling and respond to misinformation were requested by over 30 researchers and journalists.[37]

In general, there was limited information within the Initiative documents or online to enable a robust review of product uptake. It was also difficult to discern if some were still active.

Engagement with events

There is evidence that large events convened under the Initiative were widely attended and resonated with their respective audiences. Reflections on the value generated by these events are discussed in the Community building section. A selection of large events are highlighted below.

- 2017. The Global Symposium on AI & Inclusion, cohosted by **ITS Rio** and **BKC** on behalf of the Network of Centers, involved over 170 participants from more than 40 countries and took place over the course of three days.[38]
- 2017. **AI Now** Symposium had a total of 103 attendees to the Experts Workshop; its Public Symposium sold out with 299 attendees (with a waiting list of 390 people). They also had three "viewing parties" hosted at Google offices in the US and UK and a total of 2,310 viewers on the livestream.[39]
- 2018. While the **FAT ML/ACM FAccT[40]** conference did not reference uptake, one of the organizers stated that the conference"has been more successful than I ever would have imagined. It now attracts an enormous number of submissions and attendees, with many papers published at the venue going on to have a very significant impact on the research community, on practice in industry, and on policy."[41]
- 2019. AI and Digital Policy in China organized by **New America Foundation** featured panels with experts from US and international universities as well as media. They received 186 registrations.[42]
- 2021. **COGSEC** brought together 874 virtual attendees from around the world, representing a broad mix of researchers, journalists, activists, and government participants.[43]

Educational and training programs

Educational and training programs, such as university courses, seminars, and practitioner forums, supported under the Initiative broadly sought to impact the quantity and quality of scholars and practitioners working on ethics and governance of AI. Additionally, a number of grantees developed resources that targeted communities and civil society to improve awareness and promote dialogue on pertinent AI ethics and governance issues.

[35] Michael Morisy (Muckrock), online interview, December 17, 2021.
[36] Ed Bice (Meedan), online interview, November 23, 2021; Meedan, "Bot Garden User Guide," Medium, June 12, 2019, https://medium.com/meedan-user-guides/bot-garden-user-guide-adc2c1d743f3; "Check," Meedan, https://meedan.com/check.
[37] "Annual Report (March 2020–Feb 2021)," Tattle Technologies, accessed March 17, 2022, https://tattle.co.in/report/2020-report/.
[38] BKC, year one final report, November 2018; "Global Symposium on AI & Inclusion," Berkman Klein Center, https://aiandinclusion.org/#symposium.
[39] AI Now Symposium, final report, December 2017; "2017 Experts Workshop," AI Now, July 10, 2017 https://ainowinstitute.org/AI_Now_Program_2017.pdf.
[40] FAT ML evolved to ACM FAccT conference and is referred to as such in the rest if the report
[41] Cornell University FAT ML/ACM FAccT, Caribou Digital-administered online survey, December 6, 2021; "2018 Schedule," FAT/ML 2018, https://www.fatml.org/schedule/2018.
[42] New America Foundation, final report, October 2021; "AI and Digital Policy in China," Stanford University Human-Centered Artificial Intelligence, October 29, 2019, https://hai.stanford.edu/events/ai-and-digital-policy-china.
[43] Tim Hwang (previous director of the Initiative), final report, October 2021; "COGSEC 2021," University of Texas at Austin Center for Media Engagement, https://cogsec.online/.

Four examples of university courses on AI are highlighted. During their grant, **ITS Rio** developed a number of live online courses. Their course on "Artificial Intelligence and Ethics" ran twice with 190 students, and "AI, Open Data and Health" had more than 50 students. **ITS Rio** also ran a summer school with **Digital Asia Hub** on disinformation.[44] **BKC** and **Media Lab** referenced three courses on AI ethics and governance topics. These included a joint course on Applied Ethical and Governance Challenges in AI (2018/2019) alongside two seminars, "Compliance and Computation" (**BKC**, 2017) and "AI and the Law" (**BKC**, 2018).[45] These courses do not appear to be offered today. During their grant, **Princeton** set up a graduate seminar where 20 students interacted with the **Princeton Dialogues on AI and Ethics** team to test some of the ethics and governance of AI case studies they developed.[46] Ben Zevenbergen, who worked on the **Princeton** project, noted that:

> "the sessions have led to a grad seminar at Princeton, [in which] lots of students have enrolled. Several of those have changed their direction after finishing their PhDs or their masters [and] also are now working in this space. I know that the university is thinking through a formal class on this ..."[47]

Ben Zevenbergen added that there are now numerous available courses related to the ethics and governance of AI. This was validated via a cursory search; for example, Oxford University launched the Institute for Ethics in AI in 2021; the University of Cambridge launched a Master of Studies in AI Ethics and Society in 2021, which was developed and taught by the university's Leverhulme Centre for the Future of Intelligence; and the University of York now offers a master's degree in Philosophy of Artificial Intelligence.[48] There are also dozens of online courses on ethics and governance of AI offered by numerous universities, including London School of Economics, University of Helsinki, Seattle University, Université de Montréal, and University of Edinburgh.[49]

Beyond formal education, a number of other practitioner (tech and media) and public-focused training initiatives were instigated. A sample is shared below; however, insights on engagement were rarely reported.

Practitioner-focused:

- **AGTech Forum**: Launched by **BKC** in 2017, the Forum aimed to bring state attorneys general and their staff up to speed on issues related to privacy, cybersecurity, and AI. The Forum reported to have hosted representatives from 36 offices over the course of four biannual forums.[50]

Public-focused:

- **metaLAB's AI + Art**: The Initiative also supported **Harvard**'s **metaLAB's AI + Art** portfolio. They produced 10 projects, staged more than 45 exhibitions in 11 countries, were covered in over 25 articles, taught 9 workshops and courses, and gave over 50 public talks.[51]
- **Learning Experiences**: **BKC** developed a set of open educational playlists to help the public better understand AI systems and engage with their ethical challenges. These resources were released under an open-source license, and Facebook has adopted the Learning Experiences and translated them into 30+ languages.[52]

[44] ITS Rio, final report, July 2018.
[45] John DeLong, "Compliance and Computation Overview," Berkman Klein Center, https://cyber.harvard.edu/teaching/courses/2017/fall/ComplianceandComputation;
[46] Princeton CITP, interim report, January 2019; "Dialogues on AI and Ethics," Princeton University, https://aiethics.princeton.edu/.
[47] Ben Zevenbergen (Princeton CITP), online interview, December 9, 2021.
[48] "About," Leverhulme Centre for the Future of Intelligence, http://lcfi.ac.uk/about/; "MA Philosophy of Artificial Intelligence," University of York, https://www.york.ac.uk/study/postgraduate-taught/courses/ma-philosophy-artificial-intelligence/.
[49] Murat Durmus, "16 Recommended Free AI-Ethics, Data Ethics and XAI Online Courses to Get Started Right Away," *Nerd For Tech* (blog), February 2, 2022, https://medium.com/nerd-for-tech/5-recommended-free-ai-ethics-online-courses-to-get-started-right-away-2bc5daf4e417.
[50] BKC, year two final report, September 2019; "AGTech Forum," Berkman Klein Center, https://cyber.harvard.edu/research/AGTechForum.
[51] BKC, year two final report, September 2019; "AI + ART," metaLAB, https://metalabharvard.github.io/projects/aiandart/.
[52] BKC, year one final report, November 2018; "AI: Educational Activities," Berkman Klein Center, https://cyber.harvard.edu/projects/ai-educational-activities.

Reflections on centrality of Initiative assets

Behind each of the hundreds of assets developed and convenings held under the Initiative is a set of unique—though often unarticulated—ambitions regarding what would constitute successful reach and uptake. For some this may have been scale. For others, getting in front of a dozen tech or public institutions would register as a resounding success. In lieu of evaluating each of these assets on their own terms, we turn to the aggregate view:

- Citations ranged from zero to hundreds; this is to be expected with the volume of research produced.
- A selection of grantees highlighted coverage by an impressive list of top-tier and niche media outlets and reached audiences around the world.
- Product uptake was varied, with two stand-out examples of high and continued uptake from technology companies. While there was evidence of interest in other referenced products, there was limited evidence of uptake, continued deployment, or the extent to which learnings from the product demonstration were applied elsewhere.
- Insights from the events sampled suggested high relevance, engagement, and value to their target audiences.
- Insights on the educational and training programs were limited, but attendance/engagements that were reported showed demand. However, most programs (appear) to no longer be offered, and while **Princeton** cited they were developing a course, we note that there are now several universities around the world that seek to meet the demand for an education on ethics and governance of AI.

Unsurprisingly, the aggregate view offers a mixed, though mostly positive, picture regarding the centrality of the multitude of assets supported under the Initiative, indicating that many were meeting a demand in knowledge and services.

An informed public sector and private sector

> "A lot of our work ... was making big impact through ... pilot projects and concrete things [like] municipal procurement ... So it's a way of saying that our Theory of Change particularly focuses sometimes on the microcosm, because that's where the action is in this domain and where a lot of the policy decisions are being made, even though they might not feel like policy decisions."[53]

Most researchers and advocacy organizations publish insights with a minimum ambition of contributing to policy discussions. **BKC** and **Media Lab** cited an objective "to conduct evidence-based research to provide guidance to decision-makers in the private and public sectors."[54] As highlighted in Figure 1, over a quarter of grantees (26%) cited engagement with the technology industry, and more (38%) cited engagement with a multitude of policymakers and institutions in international, national, and local governments. The engagement mechanisms varied, and often multiple tactics were used in a single project, for example, scheduling one-one meetings, creating curated convenings, offering and hosting briefings, and using own networks and media coverage to get content in front of the target audience. Gordon Pennycook of the **University of Regina** cited an aim shared by a number of grantees, that the "... funding put the research on the map, which gets us in the room to talk about what we think [...] they should do."[55]

This section is concerned with the extent and depth of grantees' contributions to informing the public and private sectors on AI systems they may develop, procure, use, or regulate. Input to industry/technology companies are reviewed first, followed by a number of public sector institutions.

[53] Jonathan Zittrain (BKC), online interview, October 19, 2021.
[54] BKC and Media Lab, proposal narrative, April 12, 2017.
[55] Gordon Pennycook (University of Regina), online interview, November 17, 2021.

Technology companies

A selection of grantees provided clear examples to illustrate that industry—almost exclusively internet technology companies—were aware of and engaged with certain outputs of the Initiative.[56] The industry practices that were discussed centered on content moderation/harassment, prioritization, and accuracy.

BKC, Media Lab, and the **University of Regina** cited examples of providing insights to Facebook, via different means and for different practices. In 2018, **BKC** and **Media Lab** attended a dinner with Mark Zuckerberg to provide perspectives on moderation and AI in advance of congressional testimony.[57] **Media Lab's Gobo** project—which raised questions about content prioritization—felt that Facebook's post "Why Am I Seeing This? We Have an Answer for You" was an indication that Facebook was paying attention to the questions Gobo was asking.[58] The **University of Regina** cited they have "gotten things in front of [Facebook]" regarding improving post accuracy.[59]

Other large technology companies were reported to have engaged with the Initiative's outputs. In 2018, Jack Dorsey (cofounder of Twitter) publicly acknowledged the value of **CivilServant**'s work and allowed an independent evaluation of the effect of Twitter's harassment policies based on their research.[60] **Data & Society**'s work (presented at a conference) was cited to have influenced the thinking of a Spotify developer who "never thought that the machine learning systems that he worked on might have ripple effects that would disadvantage and even discriminate against minorities." [61] **Chequeado** also shared that, following their investigative pieces, they were contacted by technology companies (names could not be disclosed) , highlighting that they were unaware of the harm perpetrated by their systems.[62] The **University of Regina** reported that they "now have active collaborations with Google ... focused on interventions against misinformation, and are working loosely with Twitter."[63] Finally, **Mozilla**'s work on the spread of anti-vaccination conspiracy theories on social media platforms and connecting with lawmakers and journalists was reported to have resulted in (unnamed) platforms reaching out to discuss methods to avoid spreading false health information.[64]

Public sector

The bulk of evidence on informing institutions is found within the public sector. As the public sector is a broad group, this section organizes it into four thematic groups: 1) medical, 2) criminal justice, 3) international governance, and 4) national and local governance. In aggregate, a clear focal point of the Initiative was governance, particularly international and national.

Medical

Two projects—**The Data Nutrition Project** and **Media Lab's Lensing cardiolinguistics study**—provided insights to medical associations. The cardiolinguistics study, a trial to investigate supposed gender differences in angina symptoms using ML, found no differences in chest pain symptoms between genders. The findings were shared with the American Heart Association with a request to formulate new guidelines in order to correctly diagnose heart disease in women.[65] **The Data Nutrition Project**'s work with Memorial Sloan Kettering, a cancer center in New York, demonstrated

[56] Media Lab via the Moral Machine project, did work with the autonomous vehicle industry and provided examples to inform their guidelines.
[57] Media Lab, year one final report, July 2018.
[58] Media Lab, year two final report, February 2020; Anna Woorim Chung, "Gobo: Your Social Media, Your Rules," for Civic Media," June 3, 2019, https://civic.mit.edu/index.html%3Fp=2488.html; "Why Am I Seeing This? We Have an Answer for You," Facebook, March 31, 2019, https://about.fb.com/news/2019/03/why-am-i-seeing-this/.
[59] Gordon Pennycook (University of Regina), online interview, November 17, 2021; University of Regina, Caribou Digital-administered online survey, November 5, 2021.
[60] Media Lab, year one final report, July 2018.
[61] Data & Society, interim report, December 2017.
[62] Chequeado., online interview, December 22, 2021.
[63] University of Regina, Caribou Digital-administered online survey, November 5, 2021.
[64] Mozilla, final report, March 2020.
[65] Media Lab, year two final report, February 2020. Note that the cardiolinguistics study was a demonstration of AI applied for the public good, rather than *improving* AI (products/infrastructure/datasets) to better serve the public interest (reduce harms caused by AI).

that "all major data sets [...] used in dermatology that we assessed were not representative with respect to skin color."[66] Both projects provided striking findings with significant population health implications if acted upon.

Criminal justice

Three grantees shared examples of engaging and informing legislation in this area. Within the US, **BKC** and **Media Lab**, informed by the Initiative research, co-drafted a letter that urged the Massachusetts legislature to carry out a more thorough investigation of the pros and cons of pretrial risk assessment.[67] **EFF** consulted on legislation with public defenders and participated as *amicus curiae* regarding the use of algorithms in both criminal and civil contexts.[68] In the United Kingdom in 2021, researchers from **OII** were called to give evidence at the House of Lords during a public inquiry on how AI can be fairly and transparently used in the criminal justice system.[69]

International governance

Six grantees shared specific examples of informing international institutions, including the UN, the World Economic Forum (WEF), EU institutions, and international advocacy and human rights organizations. **Harvard PILAC**'s research, convenings, and tailored briefings yielded several examples of informing institutions with an international mandate. The framing of their capstone research—*Three Pathways to Secure Greater Respect for International Law concerning War Algorithm*s—was reported to have been adopted by international non-governmental organizations.[70]

BKC reported engagements that informed UN institutions. In 2019, the **BKC's Youth and Media team** launched a report which was reported to have informed UNICEF's Policy Guidance on AI for Children.[71] **BKC** also reported that they played a role in the development of a roadmap to guide UN agencies' deployment of AI in ways that advance the Sustainable Development Goals.[72] Also on the international stage, in 2018 **Access Now,** with Amnesty International, launched the Toronto Declaration to protect the rights to equality and nondiscrimination in ML systems.[73] It was endorsed by Human Rights Watch and the Wikimedia Foundation. **Access Now** credits the Declaration as key to their later success: "the promotion of it, and to use it, to convince actors is definitely one of the [...] keys to our success. [...] it was definitely part of these grants."[74]

In Europe, **Access Now** reported frequent consultations with the EU. In 2017 they participated in the EU Commission's annual review of the Privacy Shield and commented on users' rights to object to automated decision-making; they testified in the European Parliament on the proposal for a reformed e-Privacy Regulation; and when **Access Now** published their recommendations for legislators, the lead negotiators acknowledged them in a subsequent report.[75] **OII** research was cited in a number of official documents from multiple international bodies; below are four examples.[76]

1. WEF report "The Internet of Bodies is Here. This is How it Could Change Our Lives"
2. Institute of Electrical and Electronics Engineers (IEEE) Global Initiative report "Ethics of Autonomous and Intelligent Systems"

[66] Kasia Chmielinski (The Data Nutrition Project), online interview, November 29, 2021.
[67] Media Lab, year one final report, July 2018.
[68] Electronic Frontier Foundation, final report, July 2021.
[69] Sandra Wachter (Oxford Internet Institute), online interview, January 25, 2022.
[70] Dustin A. Lewis, "Three Pathways to Secure Greater Respect for International Law concerning War Algorithms, Legal Commentary," HLS PILAC, 2020, https://pilac.law.harvard.edu/three-pathways-to-secure-greater-respect-for-international-law-concerning-war-algorithms; Dustin Lewis(Harvard PILAC), online interview, November 29, 2021.
[71] BKC, year two final report, December 2019.
[72] BKC, year two interim report, January 2019. BKC referenced that they also advised entities, including the ITU's Global Symposium for Regulators and the UN's High Level Committee on Programmes, but did not provide any detail on the context.
[73] "The Toronto Declaration: Protecting the Rights to Equality and Non-discrimination in Machine Learning Systems," Access Now, May 16, 2018, https://www.accessnow.org/the-toronto-declaration-protecting-the-rights-to-equality-and-non-discrimination-in-machine-learning-systems/.
[74] Fanny Hidvegi (Access Now), online interview, January 20, 2022.
[75] Access Now, interim report, December 2017.
[76] Oxford Internet Institute, interim report, October 2020.

3. European Digital Rights report "Recommendations for a Fundamental Rights-based Artificial Intelligence Regulation: Addressing collective harms, democratic oversight and impermissible use"
4. European Parliament report "The impact of the General Data Protection Regulation (GDPR) on AI"

National and local governance

The bulk of grantees that engaged with the public sector did so at the national or, in some cases, local level. Geographically, this was with national governance institutions in the US (predominantly), mainland Europe, and the UK.

In the US, five grantees provided evidence of informing the government on topics ranging from general AI and media manipulation to Chinese cyber policy and surveillance technology use. This was active engagement via briefings, trainings, and citations of research. **New America's DigiChina** publications were cited by two major US government commissions on China—the Congressional-Executive Commission on China and the US-China Economic and Security review—and they gave two rounds of congressional testimony in 2019.[77] Dr. Joan Donovan from **Harvard University's Shorenstein Center** advised members of Congress on audio visual manipulation and disinformation and trained staffers on social media manipulation and strategies to mitigate harassment.[78] **AI Now** contributed to a Congressional Black Caucus briefing on AI.[79] **Mozilla** reported that their fellow (Camille François) testified to the House Committee on Science, Space and Technology to highlight the need for new solutions to tackle disinformation and directly answered questions from the Chair of the US Federal Elections Committee.[80] Locally, the **ACLU** testified at the Massachusetts state legislature in support of a bill that would place a moratorium on government use of face surveillance technology.[81] In Canada, **CivilServant** presented their work to a commission on freedom of expression to support their discussions with social media companies.[82]

Four grantees provided specific examples of informing national governments in Europe. In Germany, three grantees provided input. **OII**'s research was cited in documents produced by the German Data Ethics Commission.[83] **BKC** referenced their influence on Germany's AI strategy through their advisory role to the German Digital Council.[84] **Media Lab's Moral Machine** research was cited by a member of the German commission as important to the drafting of their ethical guidelines for autonomous vehicles.[85] Other national governments that were informed by the work under the Initiative include the Netherlands, via **AI Now** to discuss the implications of the Algorithm Impact Assessment (AIA),[86] and Austria and Hungary, who received **Access Now**'s comments on the "transposition" law of the GDPR.[87] Via **Access Now** and **Data & Society,** French policymakers received a consultation on their AI strategy and on AI governance in general.[88]

Perhaps due to proximity and existing relationships, **OII**'s work was referenced in several UK government documents (four examples below) and highlighted by UK Research and Innovation as a flagship case study on funding success for their 2020 AI strategy output review. This review intends to shape the government's future research funding strategy.[89]

1. UK Equality and Human Rights Commission report "Algorithms, artificial intelligence and machine learning in recruitment"

[77] New America, interim report, November 2018.
[78] Harvard Shorenstein Center, interim report, October 2019.
[79] Access Now, interim report, December 2017.
[80] Mozilla, final report, March 2020
[81] ACLU Massachusetts, Caribou Digital-administered online survey, November 1, 2021.
[82] J. Nathan Matias (CivilServant), online interview, December 20, 2021.
[83] Oxford Internet Institute, interim report, October 2020.
[84] BKC, year two interim report, January 2019.
[85] Media Lab, year two interim report, February 2020.
[86] AI Now, interim report, August 2018.
[87] Access Now, final report, July 2018
[88] Access Now, interim report, December 2017; Data & Society, interim report, December 2017.
[89] Oxford Internet Institute, interim report, October 2020.

2. UK Information Commissioner's Office report "Explaining Decisions made with AI"[90]
3. Committee on Standards in Public Life report "Artificial Intelligence and Public Standards: A Review by the Committee on Standards in Public Life"[91]
4. All Party Parliamentary Group on AI report "The rise of AI marks an opportunity for radical changes in corporate governance"

Both **GovAI** and **Data & Society**'s work was cited in the UK's "Government Response to the House of Lords Select Committee on Artificial Intelligence."[92]

Reflections on informed public and private sectors

Over one-third (n=14) of grantees provided examples of contributing to more informed public and private sectors. This equates to roughly 1 in 3 of the grantees getting the attention of various policymakers.

There were clear differences between industry and the public sector. Given there are many public sector institutions and a finite number of significant technology companies, the greater quantities of examples of informing the public sector is to be expected. Yet, in terms of the modalities, with the exception of **BKC** and **Media Lab**'s connections with the CEO of Facebook, technology companies are *less* accessible and *less* likely to reach out and share that certain insights may have been helpful.

Excluding a number of internationally mandated institutions and technology companies, the majority of examples emanated from North America, Europe, and the UK. This is broadly representative of the scope of the projects that were funded, with the exception of Latin America and Asia.

This section highlighted only explicit examples of informing specific policy discussion. It would be remiss to not acknowledge the other subtle and difficult to measure ways to inform public and private institutions, of which we have little to no information. These include the adoption of products and services like **Meedan** and **MuckRock** by media and technology companies, the multitude of conferences and events, and earned media coverage. Most of the examples here were offered or inferred, rather than explicitly measured. This means that the totality of impact on policy discussion may be greater than reported.

Changes in governance, public policy, and industry practice

An aim to directly impact policy was stated in the Initiative's foundational documents.[93] This section highlights the evidence of grantees affecting explicit changes in policy and practices of both public sector institutions and technology companies.

Technology companies

A handful of grantees could confidently say that they contributed to broad changes in the policy or practice of technology companies. One interviewee noted that it is "very hard to influence tech

[90] Information Commissioner's Office, *Explaining decisions made with AI* (Wilmslow: Information Commissioner's Office, n.d.) https://ico.org.uk/for-organisations/guide-to-data-protection/key-dp-themes/explaining-decisions-made-with-ai/.

[91] Committee on Standards in Public Life, *Artificial Intelligence and Public Standards: A Review by the Committee on Standards in Public Life* (London: Committee on Standards in Public Life, February 10, 2020) https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/868284/Web_Version_AI_and_Public_Standards.PDF.

[92] Department for Digital, Culture, Media and Sport, *Government Response to the House of Lords Select Committee on Artificial Intelligence* (London: Department for Digital, Culture, Media and Sport, February 2021), https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/963696/Government_Response_to_the_HoL_Select_Committee_on_AI_v2.pdf.

[93] Cover Memo to Ethics and Governance of AI Fund Principals, June 1, 2017.

company policy [...] that's a pretty high bar."[94] Notwithstanding, four grantees pointed to specific changes within technology companies that they felt could be attributed to their work.

Mozilla, one of the earlier grantees, noted that while direct attribution was difficult, their Fellows contributed to major platform changes, such as:[95]

- Pinterest removing anti-vaccination links
- Pinterest banning political ads
- Twitter banning political ads
- Google improving its political advertising policies
- Snap improving its political advertising policies
- Twitch improving its political advertising policies

A second example, insights from **OII**'s research—*Why fairness cannot be automated*—was adopted by HSBC and Amazon's SageMaker Clarify product.[96] For example, all Amazon Web Services customers can now use bias tests to check for and respond to discrimination in their AI systems; any technology company can use these bias tests as well as they are open access.Amazon published a write-up of the changes and directly attributed them to **OII's** research.[97] Third, the **University of Regina** is working with Jigsaw, a unit of Google that "explores threats to open societies and builds technology that inspires scalable solutions," testing the accuracy prompt on Google products.[98] The same team advised TikTok on an experiment using the accuracy prompt; however, the outcome of this work was not communicated to the university.[99] Fourth and finally, **CivilServant** reported that the comment platform Disqus added features and sent advice on harassment prevention to hundreds of thousands of websites based on their research.[100] **CivilServant** attributed their—and by extension, the **University of Regina**—influence with technology companies to the use of causal studies:

> "A hallmark of our work is that usually we're asking questions of cause and effect, which is, you know, occasionally useful for identifying problems, but also especially valuable for identifying solutions. [...] if you do this, it will reduce this problem by X or it will help contribute to a solution. And so our work on preventing harassment [...] has been adopted by many tech platforms already. And it's really easy for them to read our research and be like, oh yeah, we have a causal estimate. [...] So our work has been influential in those circles as well."[101]

Public sector

This section organizes the public sector into three thematic areas: 1) criminal justice, 2) international governance, and 3) national and local governance. The majority of the changes cited were in criminal justice and national governance; within these areas, changes centered on legal victories.

Criminal justice

Linked to their earlier work on informing the US local criminal justice institutions, **BKC**, **Media Lab**, and **EFF** cited a number of changes linked to their actions. Following **BKC** and **Media Lab**'s work urging the Massachusetts legislature to carry out a more thorough investigation of the pros and cons of pretrial assessments, the legislation was modified in line with their recommendations.[102] **EFF**

[94] Gordon Pennycook (University of Regina), online interview, November 17, 2021.

[95] Mozilla, final report, March 2020.

[96] Sandra Wachter, Brent Mittelstadt, and Chris Russell, "Why Fairness Cannot Be Automated: Bridging the Gap Between EU Non-Discrimination Law and AI," *Computer Law & Security Review* 41, no. 2021 (March 3, 2020): 105567. https://doi.org/10.2139/ssrn.3547922.

[97] Stephen Zorio, "How a Paper by Three Oxford Academics Influenced AWS Bias and Explainability Software," Amazon Science, April 1, 2021, https://www.amazon.science/latest-news/how-a-paper-by-three-oxford-academics-influenced-aws-bias-and-explainability-software.

[98] "Jigsaw," Google, https://jigsaw.google.com/.

[99] Gordon Pennycook (University of Regina), online interview, November 17, 2021.

[100] Media Lab, year one interim report, February 2018

[101] J. Nathan Matias (CivilServant), online interview, December 20, 2021.

[102] BKC, year one final report, November 2018; Media Lab, year one final report, July 2018. Joi Ito (former director of Media Lab), online interview, January 19, 2022.

supported two precedents regarding the use of AI in the criminal justice system. **EFF**'s work was cited in a legal case in the state courts (*NJ v. Pickett*), which ruled in favor of reversing the use of AI to imprison. This ruling was also reported to have influenced other courts, most notably a federal court in Pennsylvania that cited *Pickett* and agreed with arguments in **EFF**'s amicus brief (*US v. Ellis*).[103] **EFF** worked with the ACLU of Pennsylvania to file an amicus brief arguing in favor of defendants' rights to challenge DNA analysis software that implicates them in crimes. The court determined that this software company's secrecy interest could not outweigh a defendant's rights and ordered the code to be disclosed to the defense team.[104]

International governance

Six grantees cited examples of ways they informed international institutions, and two went on to cite examples of policy change that they tentatively linked to their work. **BKC** shared that the guidance provided to the AI Governance Expert Group of the Organisation for Economic Co-operation and Development (OECD) resulted in high-level AI principles which were adopted by 42 countries.[105] The ICRC's call for a ban on autonomous weapons was cited to have been influenced by Harvard **PILAC**'s research and briefings. [106]

National and local governance

Three grantees cited examples of changes in policy and practices in national and local government; these changes were exclusive to North America.[107]**AI Now** reported that, during the development of a Treasury Board Standard on Automated Decision-Making, the Government of Canada proposed that the Standard require AIA for all systems.[108] This practice is now highlighted on their "Responsible use of AI" information page.[109] **ACLU Massachusetts**'s campaign "Press Pause on Face Surveillance" was linked to several changes in policy and practice. A crowning success was their contribution to eight local bans on face surveillance in cities and towns across Massachusetts. The campaign also prompted the state legislature to commission a study on whether state law ought to impose even tighter rules on government use of the technology. In October 2021, **ACLU Massachusetts** won a landmark civil rights case at the Boston City Council, which voted unanimously to approve an ordinance that subjects AI surveillance technologies to city council oversight and accountability and public transparency.[110] **ACLU Massachusetts** linked their expanded capacity—made possible by the Initiative—to these campaign successes:

> "Thanks to your generous funding, we were able to hire Policy Counsel Emiliano Falcon-Morano. Emiliano gave us the capacity we needed on the technical and policy side to not only run a statewide public education campaign, but to fight and win seven local bans on face surveillance in cities and towns across Massachusetts."[111]

EFF was part of a multi-year campaign that called for transparency on surveillance technology; in 2020 this culminated in the passing of the Public Oversight of Surveillance Technology (POST) Act by the New York City Council.[112] This requires the NYPD to openly publish a use policy for each surveillance technology it intends to use. After this notice has been made publicly available and members of the

[103] Electronic Frontier Foundation, final report, July 2021.
[104] Electronic Frontier Foundation, final report, July 2021.
[105]BKC, year two interim report, January 2019.
[106] "Autonomous Weapons: The ICRC Recommends Adopting New Rules," International Committee of the Red Cross, August 3, 2021, https://www.icrc.org/en/document/autonomous-weapons-icrc-recommends-new-rules; Dustin Lewis (Harvard PILAC), online interview, November 29, 2021. Dustin Lewis (Harvard PILAC),online interview, November 29, 2021.
[107] Note that BKC reported that they developed the concept of information fiduciaries and reported that several (unnamed) national legislators have incorporated the concept in proposed data privacy laws. This is referenced in the impact highlights table but not called out in the main body as there is limited context.
[108] A quasi-legal instrument akin to an executive order that will bind federal departments and agencies to certain rules on how they deploy decision support systems.
[109] "Responsible Use of Artificial Intelligence (AI)," Government of Canada, https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai.html
[110] ACLU Massachusetts, Caribou Digital-administered online survey, November 1, 2021.
[111] ACLU Massachusetts, Caribou Digital-administered online survey, November 1, 2021.
[112] Nathan Sheard, "Victory! New York's City Council Passes the POST Act," Electronic Frontier Foundation, June 18, 2020, https://www.eff.org/deeplinks/2020/06/victory-new-yorks-city-council-passes-post-act.

community have had an opportunity to voice their concerns, the NYPD Commissioner must provide a final version of the surveillance impact and use policy to the City Council, the mayor, and the public.[113]

Lastly, **BKC** through their foundational work[114] on information fiduciaries–an effort to explore legal structures under which data-driven internet companies would owe a duty of loyalty to their users[115]–supported the Data Care Act proposal in 2018 and as it was reintroduced in 2021. The Data Care Act would establish a set of consumer protection duties, defined and enforced by the Federal Trade Commission, preventing technology companies from knowingly doing harm to their users. The Senator was noted to have cited BKCs work on this: "The idea has been brought up before, notably by Yale's Jack Bardin and Harvard's Jonathan Zittrain, whom Sen. Schatz has previously cited."[116]

Reflections on changes in governance, public policy, and industry practice

A quarter of the grantees (26%; n=10) linked explicit policy changes and actions to their work. This equates to roughly 1 in 4 grantees catalyzing change within the private and public sectors.

A few technology companies were shown to have adjusted a selection of practices. These included improving the quality of information on their platforms (Twitter, Pinterest, Google, Facebook); protecting the online safety of its users (Disqus); and actively assessing bias within their AI systems (Amazon and HSBC). Each of these shifts has the potential to enhance users' online (and offline) experiences. However, grantees referenced the following difficulties:

1. Grantees did not always know for certain that they impacted technology company practices. There are few incentives for technology companies to share insights that were taken on board and, reportedly, technology companies "hardly ever do [so] with anybody."[117]
2. If grantees did impact technology company practices, the ability to communicate that impact was dampened due to either NDAs or jeopardizing a continued relationship with the company.

Linked to the second point, **Meedan** suggested that policy changes may be under-reported.

> "In the case of at least a few of these partners, they take policy action based on our work. This [...] consulting has been [...] a significant aspect of our impact, [and is] somewhat underreported."[118]

Within the public sector, when change was observed, it was concentrated in the US. Many of the changes were local, such as, the ban on face surveillance technology and reassessment of AI in pretrial assessments in Massachusetts and **EFF**'s work on the POST Act in New York. **EFF**'s rulings on reversing the use of AI to implicate or imprison set a precedent that can be—and was—referred to in other campaigns. Taken together, these rulings added a level of scrutiny to ensure that software does not contribute to unjust incarceration. The impact stories within the public sector are easier to tell than those in the private sector simply because the targeted institutions are public, more accessible, and incentivised to communicate the changes they make.

Building the ethics and governance of AI community

BKC and **Media Lab** stated that the "collaboration seeks to address two of the most pressing challenges: 1) the disconnect between experts with the know-how about the design and development

[113] Electronic Frontier Foundation, final report, July 2021.
[114] The model of the Moral Machine project at the Media Lab was cited by BKC influential to their work and as one way to unpack what those new responsibilities should look like, by eliciting the expectations of real people about how they should be treated — and how they would treat others if they were in an executive role. (Jonathan Zittrain (co-founder and director of Harvard's Berkman Klein Center for Internet & Society), email to authors, 20 April, 2022.)
[115] BKC, final narrative report, May 2020.
[116] Devin Coldewey, "Senators aim to give internet companies doctor-like duties to protect our data," Techcrunch, 12 December 2018
[117] Gordon Pennycook (University of Regina), online interview, November 17, 2021.
[118] Ed Bice (Meedan), online interview, November 23, 2021.

of AI systems, and those with knowledge about the societal, ethical, and legal/policy implications of such systems and 2) information asymmetry between a relatively small group of AI experts and a very large, uninformed population that is or will be affected by these 'black box' technologies."[119] They elaborated that the collaboration would "create a common roof (in the form of community engagements, convenings, workshops, and forums) that brings together current and future thought-leaders from academia, civil society, and industry."[120] This community was intended to be inclusive of disciplines and geographies.[121]

Figure 1 shows that more than half of grantees under the Initiative (56%; n=22) included activities that meet **BKC** and **Media Lab**'s classification of community building. An addition to this classification is work involving multi- and interdisciplinary collaboration. This is implied as a central tenet of the community-building efforts, and as such certain research and pilot projects would also contribute to the envisioned community for the ethics and governance of AI. There was a valiant effort from the Initiative to implement activities that brought people and institutions together from different disciplines, domains, and, to some extent, geographies. This section describes three modalities—1) team composition, 2) convenings, and 3) supporting ethics and governance of AI institutions—in which grantees under the Initiative sought to build community and the contributions of these modalities to that effort.

Modality one: Multi- and interdisciplinary teams

The majority of projects under the Initiative had multidisciplinarity or interdisciplinarity features. The team of the Governance of Emerging Technologies (GET) Research Programme at the **OII** is comprised of lawyers, ethicists, and technologists, which was seen as very valuable in how it approaches ethics and governance of AI problems and solutions:

> "[...] truth seeking, which is very important to us, is the very academic side of things. But then we are [...] also interested to figure out what can be done, in practice [...] What could we offer in terms of finding solutions? Because it's very easy to burn a building down. It's much harder to build it up again."[122]

Another strong example of an interdisciplinary approach was the **BKC** and **Media Lab Assembly program**. The goal of this program was to combine real-world industry expertise with the socially motivated nature of academia, build lasting communities, and generate a network of socially aware AI practitioners. One reported impact of this interdisciplinary approach was an influence on participants' career trajectory. For example, a participant from the 2019 **Assembly** cohort transitioned into a full-time ethicist role at their company following the program.[123] The interdisciplinary approach was cited as critical to this impact:

> "Being part of an interdisciplinary team expanded my interest in providing a useful bridge between advancing the intellectual frontiers of how data-centric technologies impact society and translating that knowledge into action by policymakers, technologists, and civil society. This has led to an exciting new path in my career as managing director at Data & Society, working with our research and engagement teams to shift the focus onto the people most impacted by technological change."[124]

New America's work on **DigiChina** included sourcing diverse experts—on China, AI, technology, and society—to contribute to their analysis and build a sub-community of scholars. They reported contributions from 15 scholars from 13 think tanks, universities, and other organizations, highlighting that linkages between the scholars and organizations encouraged diversity and comparative inquiry.[125]

[119] BKC and Media Lab, proposal report, April 12, 2017.
[120] BKC, year one grant agreement, September 18, 2017.
[121] BKC and Media Lab, proposal, April 12, 2017.
[122] Sandra Wachter (Oxford Internet Institute), online interview, January 25, 2022.
[123] BKC, year two final report, September 2019.
[124] BKC Assembly Program and Assembly Accelerator Fund, Caribou Digital-administered online survey, quote from Ania Calderon (2019–20 Assembly Fellow), November 1, 2021.
[125] New America, interim report, November 2018.

Media Lab's VerifAI team, which explored the complexities relating to the use of AI in the criminal justice system, referred to the importance of creating space (within their project) for many different types of perspectives to inform their thinking. This was reflected in the diverse background of their team members and partner organizations. They noted that "truly multidisciplinary work is hard. It requires a real commitment from all people engaged, because you have to be willing to learn the vocabulary and conceptual frameworks of the other disciplines."[126] While most projects engaged multiple disciplines, often via events, working groups, and meetings, a select few had interdisciplinary collaboration built into their project DNA.

Modality two: Supporting new entities

Another element of community building was supporting the creation of entities aligned with the ethics and governance of AI. **BKC** reported that they played a key role in the creation of new AI-governance institutions including AI and Data Governance Center at Singapore Management University School of Law and the Thailand branch of **Digital Asia Hub (DAH)**.[127] **DAH**, based in Hong Kong, was incubated by **BKC** and aimed to provide a collaborative platform for research, knowledge sharing, and capacity building related to regional internet and society issues. **DAH**'s Executive Director Malavika Jayaram highlighted that they were one of the first in the region doing this work:

> "When we started out, we didn't really see anyone looking at trends, convergence, divergence across the region. So I think in that sense, because people saw us as someone that could help collate work and network people and connect them [...] I think they saw it as more of a collaboration [...] working with us could help elevate some of their work and connect them to [...] other organizations."[128]

The Initiative also provided seed funding to a few **Assembly** projects to continue to develop or deploy via the **Assembly Accelerator.** Seed funding was also provided to **The Markup,** a media outlet that pursues data-driven investigations into the use of technology by powerful institutions and their societal implications.

Modality three: Convenings

Across the Initiative, the number of people that attended an event, workshop, meeting, discussion forum, or policy briefing, joined a listserv, or registered as an alumni is truly likely to be in the thousands. Convening people is vital in community building—and people were convened. Sometimes this was under the broad banner of ethics and governance of AI, as was often the case for the large conferences. In other cases, convenings were held under sub-banners, such as the use of autonomous lethal weapons in conflict or media manipulation. Shared anecdotes signal the extent and depth of community building that was achieved, specifically around sharing insights, bridging disciplines and making connections:

> "I know countless stories of people who met at the [FAT ML/ACM FAccT] conference who were just chatting about shared interests and then went off and [...] wrote a paper for the next year. And more often than not the stories I hear are very interdisciplinary. So it's like not just a bunch of computer scientists who otherwise would have seen each other. [...] So it had the effect [...] of bridging some of these disciplinary divides, which is exactly what we were hoping for."[129]

The **Princeton** AI dialogue workshops were also noted to have led to collaborations between academia and industry. For example, a **Princeton** staff member was reported to have been hired by Facebook to continue exploring topics they had discussed at the workshops.[130] While the effects of these events are anecdotal, there was a "general sense" that the large conferences, such as **FAT ML / ACM FAccT**

[126] Media Lab, year one final report, July 2018.
[127] BKC, interim report, January 2019.
[128] Malavika Jayaram (Digital Asia Hub), online interview, December 16, 2021.
[129] Solon Barocas (Cornell University), online interview, December 6, 2021.
[130] Ben Zevenbergen (Princeton CITP), online interview, December 9, 2021.

"helped to establish [ethics and governance of AI] as a legitimate area of research in a number of technical fields that had previously not taken these topics very seriously."[131]

The community's internationality

Ultimately, US-based organizations were the majority of grantees, with 15% (n=6) of grantees based outside of the US (South America (2), UK (2), and Asia (2)).

A community under a common roof

The extent to which **BKC** and **Media Lab**, as the anchor institutions, achieved the objective to be the common roof of the community is not clear-cut. It is highly influenced by a limited definition of what constitutes successfully creating a common roof. In lieu of this definition, two points are reflected on: 1) leadership and 2) collaboration.

On one hand, interviewees spoke to the institutions as leaders in the field of AI ethics and governance who have contributed to the community. This view is corroborated by an article listing the top 100 academic institutions in AI, ranked by total article share in the field from 2015 to 2019. Although this evidence is dated, at that time **Harvard** was ranked first and **MIT** third.[132]

Regarding collaboration, there was some collaboration between **Media Lab** and **BKC** and some evidence of organizations funded under the Initiative crossing paths, often through attendance at each other's meetings (at the organizational rather than project level). But reports and interviews highlighted that there were no official efforts to bring Initiative projects together, an important activity in centering a community under the **BKC** and **Media Lab** "roof."

Reflections on building the ethics and governance of AI community

Evidence was forwarded from each of the three modalities—multi/interdisciplinarity, supporting new entities, and convening—deployed in the interest of community building that corroborated their value and, in the case of multi/interdisciplinary collaboration and convenings, their impact. Though the projects embodying the essence of these modalities are limited, they highlight promising practices.

Overall, interviewees expressed a variety of views on whether there was a community and the role of **BKC** and **Media Lab** in building it. Some noted that it "definitely exists" and that the two institutions "definitely contributed to it." Others felt that the Initiative "moved a community forward" via events like **FAT ML/ACM FAccT**, but that, while there is a "healthy field," the bringing together of computer science and social science "hasn't ended up with that galvanization."

Beyond hints regarding the ideal attributes of the community—cross-disciplinarity and internationality each assessed above—there is limited evidence that actions to jointly define the community were undertaken, so the concept of a community for the ethics and governance of AI remains ambiguous. Without a clear definition, it is near impossible to robustly assess if there was a Community, if it grew, or became stronger or more aligned. As one respondent noted: "the broader field of AI ethics and governance is a morass. I would have no idea how to describe or define it at this point. There is far too much going these days to be able to bring any structure to things."[133]

One thing is certain—the number of people and organizations producing outputs linked to ethics and governance of AI grew, the number of people convened under the banner of AI ethics grew and the

[131] Cornell University FAT ML/ACM FAccT, Caribou Digital-administered online survey, December 6, 2021.
[132] "Top 100 Academic Institutions in Artificial Intelligence," Nature Index, accessed February 22, 2022, https://www.natureindex.com/supplements/nature-index-2020-ai/tables/academic.

[133] Cornell University FAT ML/ACM FAccT, Caribou Digital-administered online survey, December 6, 2021.

Initiative added fuel to that growth. Insights from the **FAT ML/ACM FAccT** conference below, endorse this viewpoint, as does a sample of the Initiative's own bibliography in Annex 6:

> "[...] maintaining an up-to-date annotated bibliography seemed like a very reasonable idea at the time—one that we thought necessary for the community to grow—but this quickly became both unnecessary and practically impossible within the first year of the grant."[134]

Broad changes enabled by the Initiative

This section outlines changes outside of those described in the Theory of Change. The evaluation process relatively frequently surfaced two broader changes enabled by the Initiative on 1) individual career progression and/or change and 2) project or organization progression.

Career progression and change

As previously discussed, numerous knowledge assets, products, and services were produced through the contributions of the Initiative. The numerous publications and products produced through the Initiative occasionally contributed to their authors' or creators' career progression from AI ethics and governance research to developing policy or practices around AI ethics and governance. Some took positions in government or governance institutions, such as the US or EU government, while others took industry positions with companies like Google and Facebook.

While it is not possible to determine what impact these career changes may have on policy and practices of public and private institutions, embedding highly ethics and governance of AI-informed professionals within the public and private sector is aligned with the Initiative goals.

Project and organizational growth

Funding from the Initiative enabled some grantees to grow from a project to an organization. One interviewee shared, "the AI Ethics grant made **Tattle** an organization. We went from being an open source project to an organization. We quit our full-time jobs and we were like, okay, we're doing this."[135]

The Markup used Initiative funds as seed funding to hire staff and raise additional capital. By the end of their grant period in April 2019, they had raised US$25 million.[136] Since launching in 2020, they have published hundreds of articles "investigating how powerful institutions are using technology to change society."[137] **DigiChina** transitioned from a startup project within the **New America Foundation** to a growing program based at Stanford University, housed within the Program on Geopolitics, Technology, and Governance of the Cyber Policy Center at the Freeman Spogli Institute for International Studies. By the end of their grant period they had secured multi-year funding from the Ford Foundation.[138]

Lastly, the Initiative helped some established organizations move to the leading edge of the field of AI ethics and governance. For example, **Access Now** shared that Initiative funds put them at the forefront of civil society, leading around AI ethics and governance in Europe. They noted that they are the only digital rights non-governmental organization in the European Commission's AI Expert group.[139] These project-level transitions illustrate how the Initiative's contributions will continue to have an impact on the field in the future.

[134] Cornell University FAT ML/ACM FAccT, Caribou Digital-administered online survey, December 6, 2021.
[135] Tarumina Prabhakar (Tattle Technologies), online interview, December 22, 2021
[136] The Markup, final report, October 2019.
[137] https://themarkup.org/about.
[138] New America, final report, October 2021.
[139] Fanny Hidvegi (Access Now), online interview, January 20, 2022.

Durability of the Initiative's impact

Much of this report already explicitly or subtly spotlights the longevity of impact generated through this Initiative. For example:

- The 250+ knowledge assets (Annex 6) produced through contributions from the Initiative benefitting the AI ethics and governance community as it grows and expands. This also includes the multiple governance documents and other institutions that cite these assets.
- The masses of people and institutions educated and informed via Initiative resources, educational programs, training, public and media outreach, etc.
- The legal reforms which are not expected to be reversed in the near future.

Thus this section builds on these impacts and highlights a few other notable examples of the Initiative's sustained impact, as well as a few examples of projects or products that did not continue after the grant period.[140]

In addition to the previously mentioned examples of **Tattle**, **The Markup**, and **DigiChina**, five examples of projects sustaining impact are highlighted. The **FAT ML/ACM FAccT** conference, initially convened by **Cornell University**, shared that Initiative funding allowed them to move from a small annual workshop attached to the main machine learning conferences to a stand-alone conference affiliated with the Association for Computing Machinery.[141] It was previously cited that the **FAT ML/ACM FAccT** conference "helped to establish [AI ethics and governance] as a legitimate area of research in a number of technical fields that had previously not taken these topics very seriously" and that it remains an important venue to debate what AI ethics and governance research should look like and focus on.[142]

It was previously noted that the joint **BKC and Media Lab Assembly program** created value in bringing together engineers and product people from private industry alongside regulatory staffers and academics. Building off the Assembly program, it has now been incorporated into the new Institute for Rebooting Social Media, and is recruiting a new cohort of fellows this year. [143]

A third example is **CivilServant**, which is now a part of **Cornell University** and changed its name to the Citizens and Technology Lab.[144] **The Data Nutrition Project** successfully transitioned from an **Assembly** project to an organization that was independently funded by the Initiative and continues to run with a part-time team in 2022.[145] Finally, the **Digital Asia Hub**, incubated by BKC, remains active today.

It is important to note that while several projects continued to generate impact, some did not. This is not surprisingly for an initiative of this scale and diversity. Overall 18% (n=7) of grantees, representing 6% ($1,367,188) of total funding, did not report any impact did not report any impact beyond the development of outputs such as knowledge assets or AI prototypes.

Reflections on Initiative's aggregate and holistic impact

Over five years, the Initiative supported 39 grantees, distributed $23,325,884, and, without a doubt, generated a tremendous amount of activity on the topic of AI ethics and governance. The impact sections reviewed the Initiative's impact through the knowledge assets and products developed,

[140] Excluding from this analysis organizations that existed prior and did not have significant dependencies on funding (i.e., Data & Society, Mozilla Foundation, New America Foundation, Access Now, Meedan, MuckRock, etc.) and bounded research projects that were not expected after their research questions were addressed.
[141] In 2020, the name of this conference was changed to Association for Computing Machinery FAccT. Cornell University FAT ML, Caribou Digital-administered online survey, December 6, 2021.
[142] Cornell University FAT ML, Caribou Digital-administered online survey, December 6, 2021.
[143] Jonathan Zittrain (co-founder and director of Harvard's Berkman Klein Center for Internet & Society), email to authors, 20 April, 2022.
[144] CivilServant, final report, September 2020.
[145] Kasia Chmielinski (The Data Nutrition Project), online interview, November 29, 2021.

their uptake, the contributions of more informed public and private sectors, evidence of change in public policy and industry practice, the extent of community building, and the sustainability of the Initiative's work. The Initiative generated vast quantities of assets: over 250 publications, more than a dozen products, and countless engagements. One in three of the Initiative's grantees provided examples of how they contributed to more informed public and private sectors. One in four of the Initiative's grantees linked explicit policy changes and actions to their work. However, some research and products have yet to capture attention, policy change is a long game where connections help, and community building requires dedicated resources and commitment.

While it may be up to each of the funders to assess if the impact observed was sufficient, it is also important to consider the timeline of this Initiative. In 2016/2017, the field of ethics and governance of AI was relatively nascent, and so an element of foundation laying was required. This necessary work tends to weigh towards outputs, rather than longer-range impacts. It is exploratory, casts a wide net, and shifts tactics as context changes. Ultimately, Initiative leadership and funders should emerge with new knowledge and a clearer view on where resources should be focused next or have new learnings to apply to similar future initiatives.

Recommendations

For funders of current or future complex multi-funder/year/grantee initiatives, the following three broad recommendations are proposed:

1. **Design for diversity—in institutions, approaches, and geography—in selection processes.** Conducting informative activities—such as ecosystem scanning and surveys on priorities applied to the broader ecosystem—prior to the selection processes would be an opportunity to gain consensus on the gaps in research and practice and increase awareness of a broader range of institutions conducting relevant work. Another approach would be to embed a minimum level of diversity of institutions, countries or community interests that could enable the creation of a richer and more diverse set of implementers and impacts. Support for intra-initiative engagement—from internal Initiative newsletters, discussion forums or annual Initiative convenings—could further maximize the benefits brought by diverse grantee voices.

2. **Define the community mission to galvanize people and institutions.** Building communities is inherently difficult work, made more difficult if there is less clarity on the mission of the community. For future community-building initiatives, articulating the vision and mission, clarifying membership, and determining strategies to achieve the mission would galvanize people and institutions towards it.

3. **Design for impact measurement at the start of the initiatives**. Collaboratively design and agree on the initiative's Theory of Change (ideally) pre-implementation. This will describe how the Initiative is expected to contribute to change and in which conditions it might do so; that is, "if we do X, Y will happen because …"[146] Aligned with the Theory of Change, establish a small set of SMART (specific, measurable, achievable, relevant, and time-bound) metrics to monitor early progress and intended impact. Lastly, dedicate resources to regularly—quarterly or biannual—aggregate and review data generated by the initiative projects during implementation. On a similar cadence, engage projects to untangle the "how" and "why" of what is and is not working.

[146] Some signposts for a robust Theory of Change are: 1) Explicit: clearly articulate each predicted stage of change; 2) Context rich: exhibit a deep awareness of the operational context, which is fundamental to understanding impact and ergo designing impact research; 3) Plausible: the theory that the product could lead to the suggested outcomes is conceivable and agreed upon by expert stakeholders; 4) Testable: The theory is specific enough to enable credible testing; 5) Living: as a critical reflection tool it needs to be referred to and updated with insights as the Initiative is implemented.

In summary

At the end of the five-year, $23-million Initiative, funders are closer to understanding the extent and depth to which the Initiative achieved their aims to "deploy new prototypes, conduct research, directly impact both policy and technologies, build community, teams, and even institutions, and engage in education and outreach that meaningfully connects human values with the technical capabilities of artificial intelligence and related technologies."[147] Reflecting on the original aims cast in 2017, we summarize three key takeaways:

1. The Initiative fueled *substantial* growth in the field of AI ethics and governance and realized a number of impacts on both industry and public policy. Vast quantities of research assets were generated, and about a dozen technologies developed. Evidence suggests that, overall, these assets were highly relevant and often garnered high engagement. One in three grantees demonstrated that they informed policy, while one in four linked concrete changes in industry and public policy back to their work.

2. Views about the cohesiveness of the AI ethics and governance community varied. **BKC** and **Media Lab** supported their respective universities to deepen their own engagement on questions about, and solutions to, AI ethics and governance, supported team building across the 39 grantees, and even supported or created new entities. However, the Initiative structure and incentives did not result in sustained collaboration between **BKC** and the **Media Lab** or connection and collaboration within the Initiative. Community building is inherently difficult work and requires significant and dedicated resources to succeed.

3. Even operating in a field as complex and dynamic as AI, the Initiative was responsive to many key trends. This responsiveness was particularly notable across four efforts: 1) the embrace of the inherent interdisciplinarity of AI ethics and governance; 2) the inclusion and engagement of society through media, training, education, and creative public engagements, like art exhibitions; 3) the development and support of a counterweight to the vast industry resources and priorities on AI; and 4) the amplification of a diversity of voices in AI, of which more could have been done by the Initiative.

Donors have the opportunity—perhaps even the responsibility—to counterbalance market and geopolitical incentives in support of human-centric and ethical applications of AI. While it will not be possible for donors' funds to equal the amount spent on AI by industry or major governments, a broad mix of academic, technology, and civil society organizations will need to continue to play a key role in increasing public awareness and influencing policy.

The impacts of the Initiative on both the public and technology spaces are testaments to the unique role for skills brought by technologists, researchers, journalists, campaigners, and legal professionals, and indicates a promising path ahead.

[147] BKC-Media Lab, Cover Memo to Ethics and Governance of AI Fund Principals, June 1, 2017.

Annex 1: Evaluation framework and methods

In the first phase of this evaluation, a retrospective Theory of Change was developed (Annex 3). Prior to this evaluation, there was no explicit Theory of Change for the Initiative. A four-part evaluation framework was developed from this Theory of Change and the evaluation questions in the RFP:

- **Strength of the Theory of Change**: This theme addressed the suitability, consistency, clarity, and utility of the Initiative's Theory of Change, with special emphasis on the extent to which the activities of the implementing programs were clearly aligned with this broader Theory of Change.

- **Aggregate and holistic impact**: This theme focused on capturing and assessing the impacts of the implementing programs and lead institutions, not only in discrete/atomistic terms but also ideally in concert, toward the goals of the Initiative as a whole (as articulated in the Theory of Change).

- **Programmatic elements**: This theme explored the processes underpinning the Initiative—identification of initiative implementers, evaluation, transparency, efficiency, etc.

- **Reflections on the broader field of AI ethics and governance**: This theme explored the field's journey from "point A" in 2017 to "point B" at programs' end in 2021 and assessed the extent to which the Initiative helped drive the change between these points.

These components were explored through the following data collection methods:

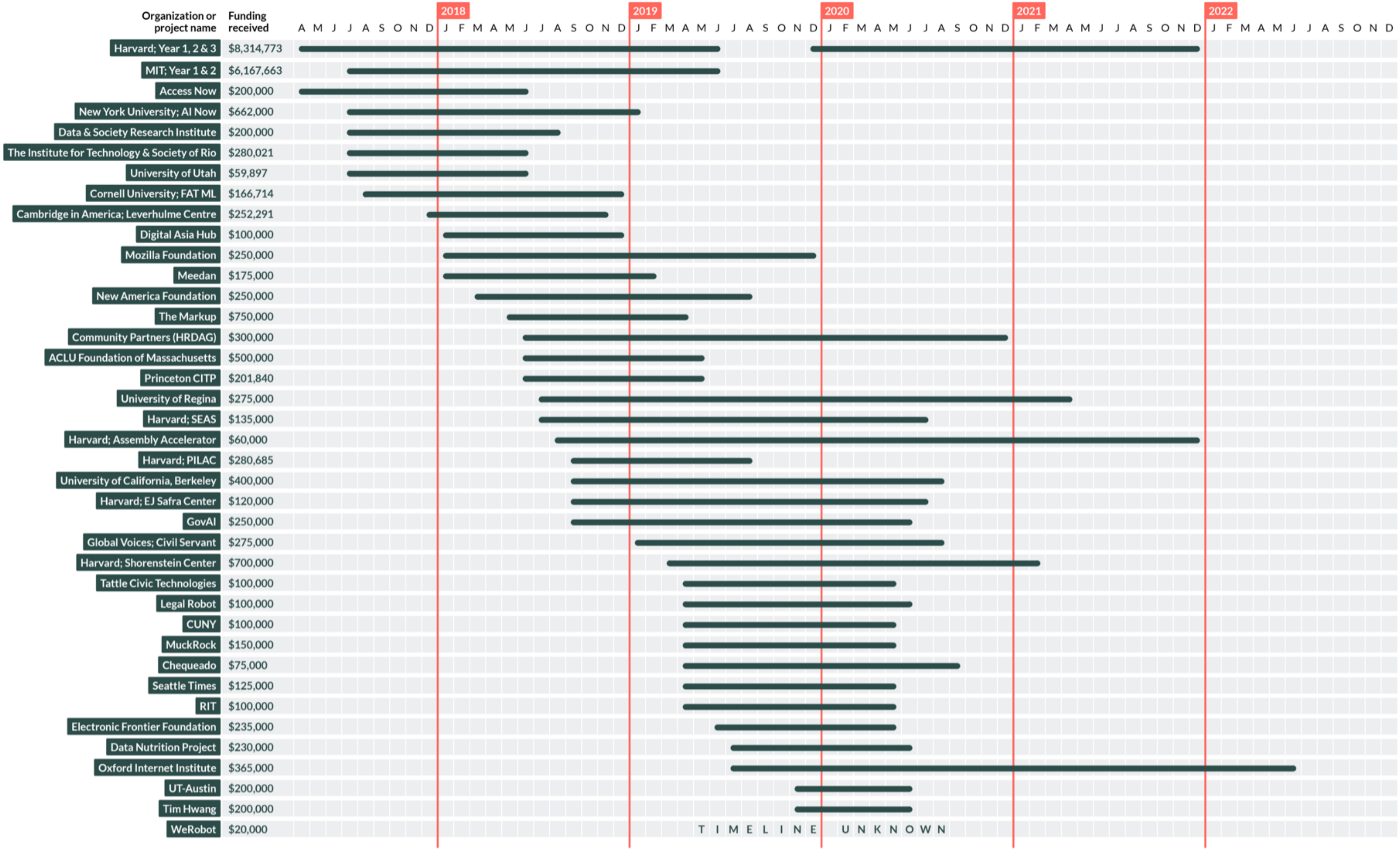
- **Document review**: To provide context to the Initiative, describe its structure, and assess its impact, we reviewed 200+ documents provided by The Miami Foundation: main proposal and grant agreements with BKC and Media Lab; grantee proposals; grantee financial documents; and grantee interim and final reports. We conducted an analytic induction of the documents using Dovetail, an online qualitative data analysis software. We started the analysis with deductive themes based on the Theory of Change and continued to look for undiscovered patterns and emergent themes throughout the analysis. We used this qualitative analysis method to ensure that we met the evaluation's purpose, while also allowing for unexpected themes to emerge.

- **Semi-structured interviews**: To help provide context to the Initiative, describe its structure, and assess its impact, we conducted semi-structured interviews over Zoom from October 2021 to February 2022. The response rate was 81% (30/37). Of the 30 completed interviews, 25 were with grantees, 2 were with outside organizations knowledgeable about the Initiative, and 3 were with BKC and Media Lab. Of the seven incomplete interviews, two did not respond to our request, two declined, and three more responded initially but did not reply to follow-up emails. We used the same data analysis method for the interviews as in the document review.

- **Surveys**: To help provide context to the Initiative, describe its structure, and assess its impact, we deployed a qualitative survey via SurveyMonkey from November to December 2021 to 95% (37/39) of grantees. At the time of survey deployment, two grantees had already been interviewed; thus, they were not sent a survey. Our response rate was 22% (8/37). We sent out two reminders to encourage organizations to respond to the survey. We used the same data analysis method for the surveys as in the interviews and document review.

Annex 2: Initiative project timeline by size

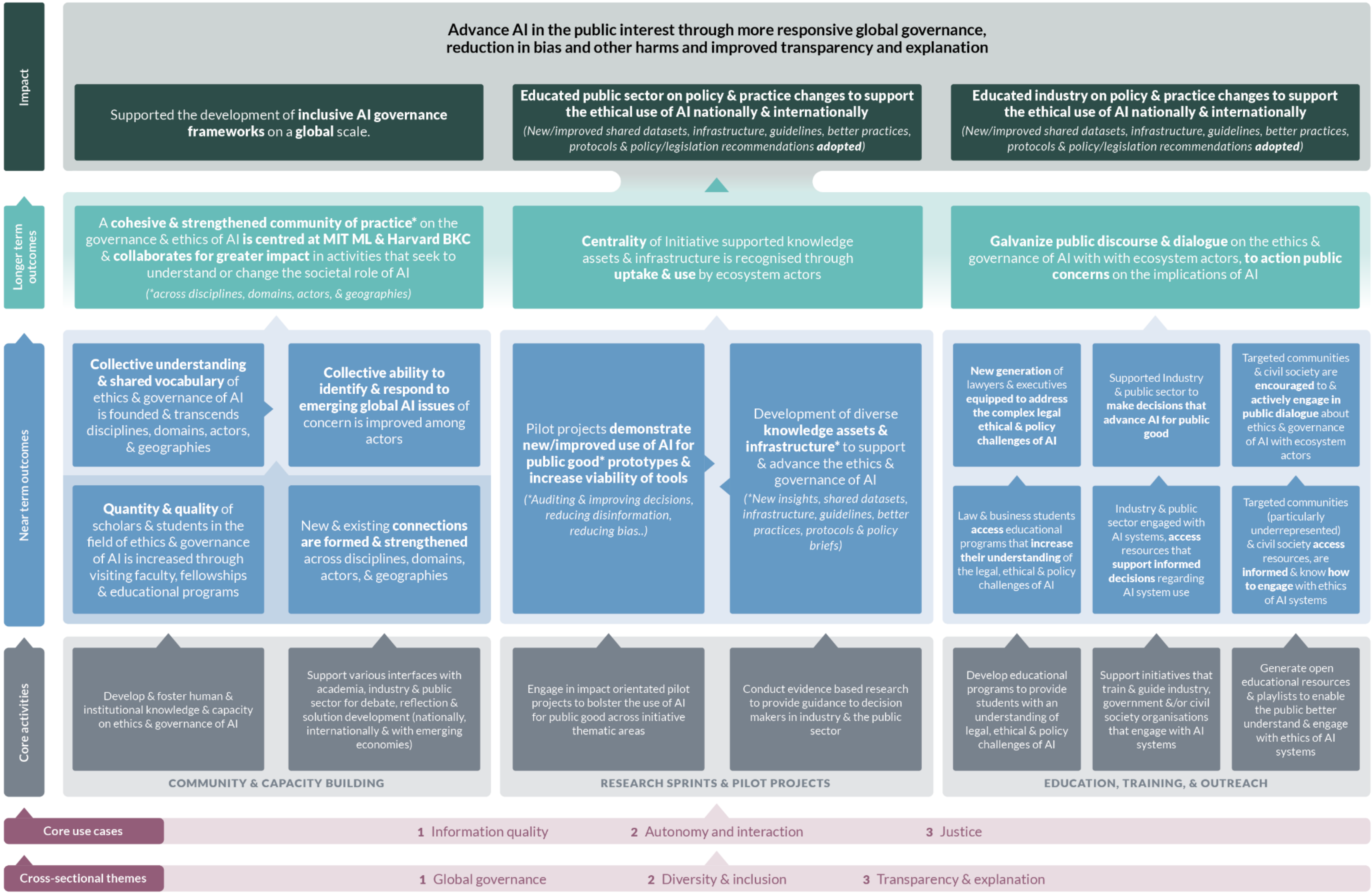
In 2017, the Initiative awarded 9 grants, with an additional 15 in 2018 and 14 in 2019. On average, 15 Initiative projects ran simultaneously each year. The year 2019 saw the highest number of projects (32), while one project finished its award in 2022. The timeline is unknown for one grantee and not included in this analysis.



Organization or project name	Funding received
Harvard; Year 1, 2 & 3	$8,314,773
MIT; Year 1 & 2	$6,167,663
Access Now	$200,000
New York University; AI Now	$662,000
Data & Society Research Institute	$200,000
The Institute for Technology & Society of Rio	$280,021
University of Utah	$59,897
Cornell University; FAT ML	$166,714
Cambridge in America; Leverhulme Centre	$252,291
Digital Asia Hub	$100,000
Mozilla Foundation	$250,000
Meedan	$175,000
New America Foundation	$250,000
The Markup	$750,000
Community Partners (HRDAG)	$300,000
ACLU Foundation of Massachusetts	$500,000
Princeton CITP	$201,840
University of Regina	$275,000
Harvard; SEAS	$135,000
Harvard; Assembly Accelerator	$60,000
Harvard; PILAC	$280,685
University of California, Berkeley	$400,000
Harvard; EJ Safra Center	$120,000
GovAI	$250,000
Global Voices; Civil Servant	$275,000
Harvard; Shorenstein Center	$700,000
Tattle Civic Technologies	$100,000
Legal Robot	$100,000
CUNY	$100,000
MuckRock	$150,000
Chequeado	$75,000
Seattle Times	$125,000
RIT	$100,000
Electronic Frontier Foundation	$235,000
Data Nutrition Project	$230,000
Oxford Internet Institute	$365,000
UT-Austin	$200,000
Tim Hwang	$200,000
WeRobot	$20,000

2018
2019
2020
2021
2022
A M J J A S O N D J F M A M J J A S O N D J F M A M J J A S O N D J F M A M J J A S O N D J F M A M J J A S O N D J F M A M J J A S O N D
TIMELINE UNKNOWN

Annex 3: Initiative Theory of Change



Impact

Advance AI in the public interest through more responsive global governance, reduction in bias and other harms and improved transparency and explanation

Supported the development of **inclusive AI governance frameworks** on a **global** scale.

Educated public sector on policy & practice changes to support the ethical use of AI nationally & internationally

*(New/improved shared datasets, infrastructure, guidelines, better practices, protocols & policy/legislation recommendations **adopted**)*

Educated industry on policy & practice changes to support the ethical use of AI nationally & internationally

*(New/improved shared datasets, infrastructure, guidelines, better practices, protocols & policy/legislation recommendations **adopted**)*

Longer term outcomes

A **cohesive & strengthened community of practice*** on the governance & ethics of AI **is centred at MIT ML & Harvard BKC** & **collaborates for greater impact** in activities that seek to understand or change the societal role of AI

*(*across disciplines, domains, actors, & geographies)*

Centrality of Initiative supported knowledge assets & infrastructure is recognised through **uptake & use** by ecosystem actors

Galvanize public discourse & dialogue on the ethics & governance of AI with with ecosystem actors, **to action public concerns** on the implications of AI

Near term outcomes

Collective understanding & shared vocabulary of ethics & governance of AI is founded & transcends disciplines, domains, actors, & geographies

Collective ability to identify & respond to emerging global AI issues of concern is improved among actors

Quantity & quality of scholars & students in the field of ethics & governance of AI is increased through visiting faculty, fellowships & educational programs

New & existing **connections are formed & strengthened** across disciplines, domains, actors, & geographies

Pilot projects **demonstrate new/improved use of AI for public good* prototypes & increase viability of tools**

*(*Auditing & improving decisions, reducing disinformation, reducing bias..)*

Development of diverse **knowledge assets & infrastructure*** to support & advance the ethics & governance of AI

*(*New insights, shared datasets, infrastructure, guidelines, better practices, protocols & policy briefs)*

New generation of lawyers & executives **equipped to address the complex legal ethical & policy challenges of AI**

Supported Industry & public sector to **make decisions that advance AI for public good**

Targeted communities & civil society are **encouraged** to & **actively engage in public dialogue** about ethics & governance of AI with ecosystem actors

Law & business students **access** educational programs that **increase their understanding** of the legal, ethical & policy challenges of AI

Industry & public sector engaged with AI systems, **access** resources that **support informed decisions** regarding AI system use

Targeted communities (particularly underrepresented) & civil society **access** resources, are **informed** & know **how to engage** with ethics of AI systems

Core activities

Develop & foster human & institutional knowledge & capacity on ethics & governance of AI

Support various interfaces with academia, industry & public sector for debate, reflection & solution development (nationally, internationally & with emerging economies)

COMMUNITY & CAPACITY BUILDING

Engage in impact orientated pilot projects to bolster the use of AI for public good across initiative thematic areas

Conduct evidence based research to provide guidance to decision makers in industry & the public sector

RESEARCH SPRINTS & PILOT PROJECTS

Develop educational programs to provide students with an understanding of legal, ethical & policy challenges of AI

Support initiatives that train & guide industry, government &/or civil society organisations that engage with AI systems

Generate open educational resources & playlists to enable the public better understand & engage with ethics of AI systems

EDUCATION, TRAINING, & OUTREACH

Core use cases: 1 Information quality; 2 Autonomy and interaction; 3 Justice

Cross-sectional themes: 1 Global governance; 2 Diversity & inclusion; 3 Transparency & explanation

Industry: those who develop & deploy AI **Public sector:** policy makers, government bodies, civil society

Ecosystem actors: Industry, government, civil society, various public bodies, government, academia, media **Bold** for emphasis on change

Annex 4: Project descriptions

Project descriptions are organized by shared activity clusters.

BKC ($8,314,773) ▼ Community building	▼ Research sprints and pilot projects	▼ Education, training, and outreach
*Multiple interdisciplinary meetings/convenings *Community spaces, e.g., ThursdAI and Global AI dialogue *Convened events, e.g., Global symposium on AI & Inclusion *Fellowships via Assembly and Techtopia	*Numerous research programs and outputs, e.g., Principled AI Project; AI and Human Rights report; Accountability of AI under the rule of law	*Various resources for education, e.g., Case study toolkit for Ethical AI; BKC policy primers on autonomous vehicles; public learning experiences *Training/outreach, e.g., public sector with AGTech Forum for attorneys general and staff; General public with metaLAB's AI + Art; *Policy engagement, e.g., advised national and international policy-makers and international bodies (ITU/UN) *Industry engagement, e.g., with Facebook on content moderation *Offered three AI ethics university courses
Media Lab ($6,167,663)		
*Multiple interdisciplinary meetings/convenings with government, academia, tech platforms, and general counsels	*Product development, e.g., BayesDB and Earshot *Applied research, e.g., Gobo Gobo; Cardio Linguistics for Atypical Angina; AIEquals *Various research programs and outputs, e.g., Society in the Loop; Scalable Cooperation Group; Moral Machine	*Training/outreach, eg., BKC/Gobo team collaboration youth digital literacy; public talks on facial recognition AI; AI for Journalism workshops *Policy engagement, e.g., on use of actuarial risk assessment; government use of facial recognition algorithm *Offered an AI ethics university course *Industry engagement, e.g., with Facebook and AV sector
NYU; AI Now ($662,000)		
*Convened the AI Now 2017 Symposium	*Research into Algorithmic Impact Assessments (AIA); AI, gender, and intersectionality	*Engagement with (Dutch) policymakers via consultations; participation in Congressional Black Caucus briefings *Developed AIA as practical framework for public agency accountability
The Institute for Technology & Society of Rio ($280,021)		
*Established working group to expand the debate on EGAI among public and private sector actors	*Developed Pegabot and Atrapabot for public to locate bots in social media	*Produced online courses *Co-taught at a summer school with Digital Asia Hub on disinformation *Trainings with Brazilian government on bots and fake news
Princeton CITP ($201,840)		

*Convened multiple transdisciplinary workshops on EGAI to develop case studies *Workshops on human rights and AI *Hosted a conference with the UN on AI and impact on the world's poorest	*Developed 6 fictional case studies on the intersection of AI and Ethics	*Developed graduate seminar on dialogues on AI and Ethics
Harvard; PILAC ($280,685)		
*Convened workshops on AI at the Frontiers of International Law concerning Armed Conflict with academia, civil society, and interactional actors	*Produced over a dozen articles and analysis pieces on AI, law, and armed conflict	*Contributed to multiple international government briefings on AI's role in conflict *Briefings with international actors, e.g., SIPRI and ICRC
Harvard; Shorenstein Center ($700,000)		
*Fellowships (×2) *Collaborations with Digital Justice lab to develop a critically engaged network of technologists on EGAI questions	*Research to develop the Media Manipulation Casebook (digital research platform)	*Consultations with tech platforms *Advising congress and training US government staffers on media manipulation and mitigation strategies *Media Manipulation Casebook as a resource for researchers, journalists, technologists, policymakers, educators, and civil society
Mozilla Foundation ($250,000)		
*Fellowship program (2× fellows)	*Research and analysis on exposing disinformation campaigns and role of AI	*Direct engagement with tech platforms, government, and journalists on misinformation including testifying to House Committee on Science, Space and Technology
Data and Society ($200,000)		
*Advisor to Partnership for AI (PAI) and participation in PAI working groups	*Variety of publications in journals and blogs on social and economic implications of AI and frameworks to assess AI	*Engagement with (EU, EUR, UK) policymakers via consultations and evidence submissions *Informal discussions with tech platforms
New America Foundation ($250,000)		
*Hosted AI and Digital Policy in China event *Hosted closed-door policy events on AI	*Produced dozens of analyses, special reports, and translations on Chinese cyberspace (including AI)	*Testified at Congress on Chinese policy on cyberspace

Digital Asia Hub ($100,000) *Convened and participated in numerous workshops, policy roundtables, and large events regionally and internationally to provide Asian perspective on EGAI	*Research via a series of deep-dive convenings on EGAI questions in developing countries	*Held a public AI art exhibition *Co-taught at a summer school with ITS Rio on disinformation
Access Now ($200,000) *Led various convenings with industry, policymakers, and civil society *Supported coordination at events, e.g., RightsCon '18; ICDPPC '17; AI summit at MWC '18	*Numerous commentaries and analyses on topics such as GDPR and e-privacy	*Engagement with (EU, EUR) policymakers via consultations *Developed policy guides for law-makers and user guides to EU data protection | Toronto deceleration
Meedan ($175,000) *Initiated various Credibility Coalition working groups	*Public release of dataset (of articles annotated from their credibility) and results papers *Developed Bot Garden as part of Check offering to automate and scale fact-checking for newsrooms	*Content moderation partnerships with tech platforms *Fact-checking partnerships with newsrooms and civil society
The Markup ($750,000) *Seed funding for new investigative data-journalism organization focused on societal effects of AI and other technologies	*Received seed funding for set up; reports were generated after the grant period	*Public awareness generated via open access reporting

Electronic Frontier Foundation ($235,000) ▼ **Research sprints and pilot projects**	▼ **Education, training, and outreach**
*Research to develop various amicus briefs on software code transparency (used in criminal justice system)	*Policy engagement via participating as amici, consulting with public defenders, and issuing amicus briefs
ACLU Foundation of Massachusetts ($500,000) *Research to inform public campaigns on government use of face surveillance technology	*Multiple practical resources for public awareness during the "press pause on face surveillance" campaign also engage in a week of action with youth action and media coverage *Policy engagement via public records lawsuit, public record requests, testifying

Chequeado ($75,000)	
*Eight investigative journalism pieces on AI and disability, work, education, justice, and health	*Guide for journalists covering AI *Media alliances to share results of research
Seattle Times ($125,000)	
*Scores of articles produced during a year-long spotlight on a range of AI topics	*Public engagement via own media channels and other media outlets *Student outreach via a tour and discussion on journalism covering AI
University of Regina ($275,000)	
*Multiple causal reasoning experiments and research published on misinformation in social media and approaches to reduce the spread of misinformation	*Engagement with industry via consultation with Facebook, Twitter, and partnerships, e.g., with Jigsaw *Public outreach via public media (*NY Times* articles) and public lectures
Global Voices; CivilServant ($275,000)	
*Developed three products for conducting research, e.g., to inform users about their data use, Reddit monitoring, privacy, and ethics software. *Multiple citizen science research outputs including methodologies	*Presented evidence to government commissions (Canada) *Publish community debriefing post=-research guided by community collaborators *Engage relevant communities in formulating research questions (community research summits) *Engaging policy via findings submitted as advice (UK Parliament) *Engagement with tech platforms on findings
Oxford Internet Institute ($365,000)	
*Development of the Governance of Emerging Technologies (GET) research program; dozens of EGAI publications generated including experiments	*Engagement with (UK) policymakers via consultations and evidence submissions on use of AI in the criminal justice system *Engagement with tech platforms and private organizations using AI *Public documentaries on research (BBC, Wired)
Tattle Civic Technologies ($100,000)	
*Developed open source tools to assist fact-checking, e.g., WhatsApp Archiver (scraper) and Tattle Khoj (dataset) *Documenting ethical considerations in archiving and opening data from Chat Apps	*Engagement with civil society via advising journalists (ad hoc) *Created a portal for journalists to access all resources
Rochester Institute of Technology (RIT) ($100,000)	

*Developed DeFake a ML tool to identify if audio and video evidence is fake *Published research describing the model	*Engagement with journalists (as testers) *Engagement with industry via ethics discussions with developers of deep fakes
Cambridge in America; Leverhulme Center ($252,291) ▼ **Community building** *Organize multiple interdisciplinary workshops to explore algorithmic interpretability requirements in light of GDPR	▼ **Research sprints and pilot projects** *Research published on methods for discovering interpretable representations in AI settings
Harvard; Assembly Accelerator ($60,000) *Fostered a community/fellowship (annual cohort) spanning sectors and disciplines to understand and develop solutions to challenges in EGAI	*Assembly cohorts developed ten frameworks, tools, and other solutions to address EGAI challenges
Harvard; EJ Safra Center ($120,000) *Fellowship program	*Published on developing oversight mechanisms to govern the development of emerging technologies
CUNY ($100,000) ▼ **Community building** *AI monitor journalist fellowship program	▼ **Education, training and outreach** *Training community and ethnic media journalists to uncover and analyze AI systems *Developed resources for journalists (expert videos, primers)
Community Partners; HRDAG ($300,000) ▼ **Research sprints and pilot projects** *Published research on AI-generated pretrial risk assessments *Collaborated on other research—using ML to target FOIA requests on police complaints, ML to predict hidden graves in Mexico—but no research outputs reported	
University of California Berkeley ($400,000) *Published research on how to better align algorithmic decision-making systems with the social desiderata of fairness and explainability *Developed WhyNot, a Python package that provides an experimental sandbox for decisions in dynamics, connecting tools from causal inference and reinforcement learning with challenging dynamic environments	
GovAI ($250,000)	

*Multiple publications on AI experts' opinions and forecasts on ethics and governance
*A public opinion survey on AI was funded but, not yet published

Harvard; SEAS ($135,000)

*Research on the impact of exposing the key features of a pre-trial assessment on judge decisions/next steps (no publication as still in analysis)

Data Nutrition Project ($230,000

*Developed tool to assess the accuracy and fairness of AI via improving datasets
*Published data nutrition framework

MuckRock ($150,000)

*Launched Sidekick, a crowdsourcing toolset of machine learning classifiers to analyze documents/datasets

Legal Robot ($100,000)

*Launched Legal Data Commons
*Further development of Legal Robot's AI for acquisition of vast quantities of public contracts and automated document analysis

Cornell University; FAT ML ($166,714)

▾ Community building

*Established a cross-disciplinary annual conference on issues of fairness, accountability, and transparency in computational systems

UT-Austin; COGSEC & Tim Hwang; COGSEC
($400,000 each)

*Convened a virtual event focused on teaching practical skills in investigating, media manipulation, and disinformation efforts

WeRobot ($20,000)

**No reports required by the grant*

Annex 5: Products developed or expanded under the Initiative

A number of products do not have specific names. This may be an omission in reporting or because the product is unnamed.

▾ Product name	*Organization name*	Description
AI benchmark repository	*University of Utah*	Algorithms, datasets, and metrics to evaluate different fairness-aware tools
AI Compass	*BKC*	Interactive dashboard that highlights connections across selected focus areas (AI+Governance, AI+Inclusion, AI+Art & Design, AI+Youth)
Atrapabot	*ITS Rio*	Tool that creates more transparency about bot usage in Mexico
Bartleby	*Global Voices (CivilServant)*	Software for large-scale management of GDPR, CCPA, and university research ethics procedures
BayesDB	*Media Lab*	Open-source AI software that enables programmers to answer data analysis questions in seconds
Bot Garden	*Meedan*	Supports fact-checking for researchers and media outlets at scale
Data Nutrition Tool	*Data Nutrition Project*	Creates a standard label for interrogating datasets
DeFake	*RIT*	Tool to detect manipulated videos
Earshot	*Media Lab*	Search tool for American talk radio
Gobo	*Media Lab*	Tool to control algorithms on social media platforms
Intelligent web scraper	*Legal Robot*	Tool to scrape public records from US local government offices
PegaBot	*ITS Rio*	Tool that creates more transparency about bot usage in Brazil
Risk Assessment Tool Database	*BKC*	Collects public information about design and implementation of risk assessment instruments used in US criminal justice system
SepsisWatch	*Data & Society*	iPad app that uses deep learning to display a patient's risk for developing sepsis
Sidekick	*MuckRock*	Crowdsourcing/API for Freedom of Information Act (FOIA) requests
FactCheck Article Scraper	*Tattle Civic Technologies*	Repository that contains a collection of scripts to scrape the fact checking sections of certain websites in India
Jod Bot	*Tattle Civic Technologies*	Telegram interface to upload media to tattle's archive and search through it.
Kosh	*Tattle Civic Technologies*	Searchable archive of multimedia content relevant to misinformation and social media in India
WhatsApp Archiver	*Tattle Civic Technologies*	Tool that consolidates chat files exported from different WhatsApp conversations into one database.
Turing Box	*Media Lab*	Two-sided marketplace that allows AI contributors to upload existing and novel algorithms for scientific study
WhyNot	*University of California Berkeley*	Python package that provides an experimental sandbox for decisions in dynamics, connecting tools from causal inference and reinforcement learning with challenging dynamic environments

Annex 6: Knowledge assets produced under the AI Initiative

The bibliography below is a sample of 256 knowledge assets produced by 30 grantees. Please note that this list is not exhaustive.

"A Compilation of Materials Apparently Reflective of States' Views on International Legal Issues Pertaining to the Use of Algorithmic and Data-Reliant Socio-Technical Systems in Armed Conflict." Cambridge, MA: Harvard Law School Program on International Law and Armed Conflict. Accessed March 3, 2022. https://pilac.law.harvard.edu/a-compilation-of-materials-apparently-reflective-of-states-views-on-international-legal-issues-pertaining-to-the-use-of-algorithmic-and-data-reliant-socio-technical-systems-in-armed-conflict.

Access Now. *A User Guide to Data Protection in the European Union: Your Rights & How to Exercise Them.* Brussels: Access Now, 2018. https://www.accessnow.org/cms/assets/uploads/2018/07/GDPR-User-Guide_digital.pdf.

Acker, Amelia, and Joan Donovan. "Data Craft: A Theory/Methods Package for Critical Internet Studies." *Information, Communication & Society* 22, no. 11 (2019): 1590–609. https://doi.org/10.1080/1369118X.2019.1645194.

ACLU of Massachusetts. "An Act to Regulate Face Surveillance," 2019. https://www.aclum.org/sites/default/files/field_documents/regulate_face_surveillance_fact_sheet_final.pdf.

———. "Face Surveillance 101," 2019. https://www.aclum.org/sites/default/files/field_documents/facesurveillance101_2.0.pdf.

———. "Face Surveillance and Government Intrusion," 2019. https://www.aclum.org/sites/default/files/field_documents/face_surveillance_and_government_intrusion.pdf.

———. "Face Surveillance and Racial Bias," 2019. https://www.aclum.org/sites/default/files/field_documents/racial_bias_and_fs.pdf.

———. "Why We Need Stronger Legal Protections," 2019. https://www.aclum.org/sites/default/files/why_regulations.pdf.

Acosta, Aida Joaquin. "3 Practical Tools to Help Regulators Develop Better Laws and Policies." Berkman Klein Center for Internet & Society at Harvard University, July 2018. https://cyber.harvard.edu/sites/default/files/2018-07/2018-07_AVs04_1.pdf.

———. "5 Technological Factors Regulators and Policymakers Need to Know." Berkman Klein Center for Internet & Society at Harvard University, July 2018. https://cyber.harvard.edu/sites/default/files/2018-07/2018-07_AVs01_1.pdf.

———. "24 Essentials of a SWOT Analysis Policymakers Need to Consider." Berkman Klein Center for Internet & Society at Harvard University, July 2018. https://cyber.harvard.edu/sites/default/files/2018-07/2018-07_AVs02_0.pdf.

———. "What Governments Across the Globe Are Doing to Seize the Benefits of Autonomous Vehicles." Berkman Klein Center for Internet & Society at Harvard University, July 2018. https://cyber.harvard.edu/sites/default/files/2018-07/2018-07_AVs03_0.pdf.

Adel, Tameem, Isabel Valera, Zoubin Ghahramani, and Adrian Weller. "One-Network Adversarial Fairness." *Proceedings of the AAAI Conference on Artificial Intelligence*, 33, no. 1 (2019): 2412–20. https://doi.org/10.1609/aaai.v33i01.33012412.

Adel, Tameem, Zoubin Ghahramani, and Adrian Weller. "Discovering Interpretable Representations for Both Deep Generative and Discriminative Models." *Proceedings of the 35th International Conference on Machine Learning, PMLR* 80 (2018): 50–59. https://proceedings.mlr.press/v80/adel18a.html.

Adjodah, Dhaval, Tim Klinger, and Joshua Joseph. "Symbolic Relation Networks for Reinforcement Learning." *32nd Conference on Neural Information Processing Systems* (2018). https://r2learning.github.io/assets/papers/CameraReadySubmission%203.pdf.

AI & Inclusion Symposium. "Resources." *AI & Inclusion Symposium* (blog), September 17, 2017. https://aiandinclusionsymposium.com/inputs/.

Alabdulkareem, Ahmad, Morgan R. Frank, Lijun Sun, Bedoor AlShebli, César Hidalgo, and Iyad Rahwan. "Unpacking the Polarization of Workplace Skills." *Science Advances* 4, no. 7 (2018): eaao6030. https://doi.org/10.1126/sciadv.aao6030.

Alade, Yewande, Christine Kaeser-Chen, Elizabeth Dubois, Chintan Parmar, and Friederike Schüür. "Towards Better Classification." MIT Media Lab. https://kaleidoscope.media.mit.edu/white-paper.

Amnesty International and Access Now. "The Toronto Declaration." Toronto Declaration, May 2018. https://www.torontodeclaration.org/declaration-text/english/.

Angehrn, Zuzanna, Clementine Nordon, Andrew Turner, Dianne Gove, Helene Karcher, Alexander Keenan, Monika Neumann, Jelena Sostar, and Frederic de Reydet de Vulpillieres. "Ethical and Social Implications of Using Predictive Modeling for Alzheimer's Disease Prevention: A Systematic Literature Review Protocol." *BMJ Open* 9, no. 3 (March 1, 2019): e026468. https://doi.org/10.1136/bmjopen-2018-026468.

Ávila-Zesatti, Cristina. "Un mapa de datos para predecir la existencia de fosas clandestinas en México." *Chequeado* (blog), September 30, 2020. https://chequeado.com/investigaciones/un-mapa-de-datos-para-predecir-la-existencia-de-fosas-clandestinas-en-mexico/.

Awad, Edmond, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. "The Moral Machine Experiment." *Nature* 563, no. 7729 (November 2018): 59–64. https://doi.org/10.1038/s41586-018-0637-6.

Balkin, Jack M., and Jonathan Zittrain. "A Grand Bargain to Make Tech Companies Trustworthy." The Atlantic, October 3, 2016. https://www.theatlantic.com/technology/archive/2016/10/information-fiduciary/502346/.

Balmaceda, Tomás. "Inteligencia artificial y discapacidad: cuando los algoritmos son herramientas de exclusión." *Chequeado* (blog), September 7, 2020. https://chequeado.com/investigaciones/inteligencia-artificial-y-discapacidad-cuando-los-algoritmos-son-herramientas-de-exclusion/.

Barabas, Chelsea. "Beyond Bias: 'Ethical AI' in Criminal Law." In *The Oxford Handbook of Ethics of AI*, edited by Markus D. Dubber, Frank Pasquale, and Sunit Das. Oxford: Oxford University Press, 2020. https://doi.org/10.1093/oxfordhb/9780190067397.013.47.

Barabas, Chelsea, Madars Virza, Karthik Dinakar, Joichi Ito, and Jonathan Zittrain. "Interventions over Predictions: Reframing the Ethical Debate for Actuarial Risk Assessment." *Proceedings of the 1st Conference on Fairness, Accountability and Transparency, PMLR* 81 (2018): 62–76. https://proceedings.mlr.press/v81/barabas18a.html.

Barabas, Chelsea, Ruha Benjamin, John Bowers, Meredith Broussard, Joy Buolamwini, Sasha Constanza-Chock, Kate Crawford, et al. "Technical Flaws of Pretrial Risk Assessments Raise Grave Concerns." https://dam-prod.media.mit.edu/x/2019/07/16/TechnicalFlawsOfPretrial_ML%20site.pdf.

Basl, John, and Jeff Behrends. "Why Everyone Has It Wrong About the Ethics of Autonomous Vehicles." In *Frontiers of Engineering: Reports on Leading-Edge Engineering from the 2019 Symposium*. Washington, DC: The National Academies Press, 2020. https://doi.org/10.17226/25620.

Basl, John, and Joseph Bowen. "AI as a Moral Right-Holder." In *The Oxford Handbook of Ethics of AI*, edited by Markus D. Dubber, Frank Pasquale, and Sunit Das. Oxford: Oxford University Press, 2020.

Bavitz, Christopher, Sam Bookman, Jonathan Eubank, Kira Hessekiel, and Vivek Krishnamurthy. "Assessing the Assessments: Lessons from Early State Experiences In the Procurement and Implementation of Risk Assessment Tools." Berkman Klein Center Research Publication No. 2018-8, Social Science Research Network, Rochester, NY, December 1, 2018. https://doi.org/10.2139/ssrn.3297135.

Behrends, Jeff, and John Basl. "Trolleys and Autonomous Vehicles: New Foundations for the Ethics of Machine Learning." In *Autonomous Vehicle Ethics: Beyond the Trolley Problem*. Oxford: Oxford University Press, forthcoming.

Benkler, Yochai, Robert Faris, and Hal Roberts. *Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics*. New York: Oxford University Press, 2018. https://doi.org/10.1093/oso/9780190923624.001.0001.

Bollier, David. *Artificial Intelligence, the Great Disruptor: Coming to Terms with AI-Driven Markets, Governance and Life*. Washington, DC: Aspen Institute, 2018. https://csreports.aspeninstitute.org/documents/AI2017.pdf.

Bond, Raymond R., Maurice Mulvenna, and Hui Wang. "Human Centered Artificial Intelligence: Weaving UX into Algorithmic Decision Making." In *RoCHI 2019*, 8, 2019. http://rochi.utcluj.ro/articole/7/RoCHI2019-Bond.pdf.

Boyd, Danah, and M. C. Elish. "Don't Believe Every AI You See." *New America* (blog), November 26, 2018. http://newamerica.org/pit/blog/dont-believe-every-ai-you-see/.

Bridgeman, Tess. "The Viability of Data-Reliant Predictive Systems in Armed Conflict Detention." *Humanitarian Law & Policy Blog* (blog), April 8, 2019. https://blogs.icrc.org/law-and-policy/2019/04/08/viability-data-reliant-predictive-systems-armed-conflict-detention/.

Cáceres, Romina, and Jazmín Acuña. "Se busca empleador: inteligencia artificial para conseguirte un trabajo." *Chequeado* (blog), September 17, 2020. https://chequeado.com/investigaciones/se-busca-empleador-inteligencia-artificial-para-conseguirte-un-trabajo/.

Campolo, Alex, Madelyn Sanfilippo, Meredith Whittaker, and Kate Crawford. "AI Now 2017 Report." AI Now, 2017. https://ainowinstitute.org/AI_Now_2017_Report.pdf.

Center, Berkman Klein. "Data Commons Version 1.0: A Framework to Build Toward AI for Good." *Berkman Klein Center Collection* (blog), June 22, 2018. https://medium.com/berkman-klein-center/data-commons-version-1-0-a-framework-to-build-toward-ai-for-good-73414d7e72be.

Chertoff, Phillip, Jeff Fossett, Saffron Huang, Yonadav Shavit, and Irene Solaiman. "Municipal Government Automated Decision-Making Systems Playbook." Harvard University. Accessed March 1, 2022. https://drive.google.com/file/d/1t_rgn3p2gLmOZhwUrYFFLFn69l2-yXU2/view?usp=embed_facebook.

Chintha, Akash, Bao Thai, Saniat Javid Sohrawardi, Kartavya Bhatt, Andrea Hickerson, Matthew Wright, and Raymond Ptucha. "Recurrent Convolutional Structures for Audio Spoof and Video Deepfake Detection." *IEEE Journal of Selected Topics in Signal Processing* 14, no. 5 (August 2020): 1024–37. https://doi.org/10.1109/JSTSP.2020.2999185.

Chmielinski, Kasia S, Sarah Newman, Matt Taylor, Josh Joseph, Kemi Thomas, Jessica Yurkofsky, and Yue Chelsea Qiu. "The Dataset Nutrition Label (2nd Gen): Leveraging Context to Mitigate Harms in Artificial Intelligence." *NeurIPS 2020 Workshop on Dataset Curation and Security*, 2020, 7. http://securedata.lol/camera_ready/26.pdf.

Chung, Anna Woorim. "Gobo: Your Social Media, Your Rules – MIT Center for Civic Media," June 3, 2019. https://civic.mit.edu/2019/06/03/gobo-your-social-media-your-rules/../../../../index.html%3Fp=2488.html.

———. "How Automated Tools Discriminate Against Black Language – MIT Center for Civic Media," January 24, 2019. https://civic.mit.edu/2019/01/24/how-automated-tools-discriminate-against-black-language/../../../../index.html%3Fp=2402.html.

Citizens and Technology Lab at Cornell University. "COVID-19 Algorithms and Public Health." *Citizens and Technology Lab* (blog). Accessed March 3, 2022. https://citizensandtech.org/covid-19-algorithms-and-public-health/.

———. "The Effects of Saying Thanks Online." *Citizens and Technology Lab* (blog). Accessed March 3, 2022. https://citizensandtech.org/research/how-do-wikipedians-thank-each-other/.

COGSEC Conference. *Aric Toler | Media Forensics in the Field*, 2021. https://www.youtube.com/watch?v=ZVpOOxZHNVc.

———. *Craig Silverman | Connecting the Dots: A Primer on Uncovering Fraudulent Networks Online*, 2021. https://www.youtube.com/watch?v=xUdX_BxxrxY.

———. *Elisabeth Bik | Spotty Science: The Art and Techniques of Debunking Scientific Reports*, 2021. https://www.youtube.com/watch?v=k6N9zT19Pis.

———. *Emily Gorcenski | Data and Daylight: New Tools for Exposing and Countering Neofascist Actors*, 2021. https://www.youtube.com/watch?v=E6iIVEW-Qjg.

———. *Fireside Chat with Dan Rather, Natalie Wynn (ContraPoints), and Avery Trufelman*, 2021. https://www.youtube.com/watch?v=fLKYp0LyjMk.

———. *Joan Donovan | Deplatforming: Case Studies from the Field*, 2021. https://www.youtube.com/watch?v=62t8lO00v8g.

———. *Max Weiss | The Threat of Generative Text in Public Solicitation: Vulnerabilities and Solutions*, 2021. https://www.youtube.com/watch?v=bPFY6oXQ3ho.

———. *Praveen Sinha | You've Been Doxxed, Now What? (Part 1)*, 2021. https://www.youtube.com/watch?v=Q0wBJXQA7Rk.

———. *Praveen Sinha | You've Been Doxxed, Now What? (Part 2)*, 2021. https://www.youtube.com/watch?v=kmSMYeKpJU8.

———. *Talia Lavin | Into the Lion's Den: How to Infiltrate Extremist Spaces*, 2021. https://www.youtube.com/watch?v=KUYwrPHb3ak.

———. *Yudhanjaya Wijeratne | Building an Effective Fact-Checking Operation on the Cheap*, 2021. https://www.youtube.com/watch?v=J1TsYWUofGs.

Coplin, Abigail, Paul Triolo, Elsa Kania, Rui Zhong, Benjamin Larsen, and John Lee. "Experts: Xi's Science and Technology Speech Echoes and Updates Deng Xiaoping." *DigiChina* (blog). Accessed March 2, 2022. https://digichina.stanford.edu/work/experts-xis-science-and-technology-speech-echoes-and-updates-deng-xiaoping/.

Costigan, Johanna. "Four Specialists Describe Their Diverse Approaches to China's AI Development." *DigiChina* (blog). Accessed March 2, 2022. https://digichina.stanford.edu/work/four-specialists-describe-their-diverse-approaches-to-chinas-ai-development/.

Dai, Jessica, Sina Fazelpour, and Zachary Lipton. "Fair Machine Learning Under Partial Compliance." In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 55–65. New York, NY, USA: Association for Computing Machinery, 2021. https://doi.org/10.1145/3461702.3462521.

de Andrade, Eduardo Goulart. "Covid-19 y tecnología: Brasil utiliza inteligencia artificial y aplicación de seguimiento para tratar de combatir la pandemia." *Chequeado* (blog), September 30, 2020. https://chequeado.com/investigaciones/covid-19-y-tecnologia-brasil-utiliza-inteligencia-artificial-y-aplicacion-de-seguimiento-para-tratar-de-combatir-la-pandemia/.

Deeks, Ashley. "Detaining by Algorithm." *Humanitarian Law & Policy Blog* (blog), March 25, 2019. https://blogs.icrc.org/law-and-policy/2019/03/25/detaining-by-algorithm/.

DigiChina, Stanford Program on Geopolitics, Technology, and Governance, and New America. "AI Policy and China: Realities of State-Led Development." Stanford-New America DigiChina Project, October 29, 2019. https://d1y8sb8igg2f8e.cloudfront.net/documents/DigiChina-AI-report-20191029.pdf.

DiResta, Renee. "The Complexity of Simply Searching For Medical Advice." *Wired*. Accessed March 2, 2022. https://www.wired.com/story/the-complexity-of-simply-searching-for-medical-advice/.

DiResta, Renee, and Shelby Grossman. "Potemkin Pages & Personas: Assessing GRU Online Operations, 2014-2019," November 12, 2019. https://cyber.fsi.stanford.edu/publication/potemkin-think-tanks.

Donovan, Joan. "First They Came for the Black Feminists." *The New York Times*, August 15, 2019. https://www.nytimes.com/interactive/2019/08/15/opinion/gamergate-twitter.html.

———. "How Hate Groups' Secret Sound System Works." *The Atlantic*, March 17, 2019. https://www.theatlantic.com/ideas/archive/2019/03/extremists-understand-what-tech-platforms-have-built/585136/.

———. "How News Organizations Should Cover White Supremacist Shootings, According to a Media Expert." PBS NewsHour Weekend, August 4, 2019. https://www.pbs.org/newshour/show/how-media-coverage-contributes-to-white-supremacist-rhetoric.

———. "Opinion | Men like the El Paso Shooter Aren't 'Lone Wolves' — They're Never Alone Online." *NBC News | Think* (blog), August 5, 2019. https://www.nbcnews.com/think/opinion/el-paso-shooter-wasn-t-lone-wolf-his-so-called-ncna1039201.

———. *PBS NewsHour | How Media Coverage Contributes to White Supremacist Rhetoric | Season 2019*, 2019. https://www.pbs.org/video/how-media-coverage-contributes-to-white-supremacist-rhetoric-1564946365/.

———. "Perspective | How Trump Put Himself in Charge of Twitter's Decency Standards." *Washington Post*, July 19, 2019. https://www.washingtonpost.com/outlook/2019/07/19/how-trump-put-himself-charge-twitters-decency-standards/.

Donovan, Joan, and Brian Friedberg. "Source Hacking: Media Manipulation in Practice." Data & Society Research Institute, September 4, 2019. https://datasociety.net/wp-content/uploads/2019/09/Source-Hacking_Hi-res.pdf.

Doshi-Velez, Finale, Mason Kortz, Ryan Budish, Christopher Bavitz, Samuel J. Gershman, David O'Brien, Kate Scott, et al. "Accountability of AI Under the Law: The Role of Explanation." SSRN Scholarly Paper. Rochester, NY: Social Science Research Network, November 3, 2017. https://doi.org/10.2139/ssrn.3064761.

Doshi-Velez, Finale, Mason Kortz, Ryan Budish, Christopher Bavitz, Samuel J. Gershman, David O'Brien, Stuart Shieber, Jim Waldo, David Weinberger, and Alexandra Wood. "Accountability of AI Under the Law: The Role of Explanation." *SSRN Electronic Journal*, 2017. https://doi.org/10.2139/ssrn.3064761.

Elish, M C. "(Dis)Placed Workers: A Study in the Disruptive Potential of Robotics and AI," 37, 2018. https://conferences.law.stanford.edu/werobot/wp-content/uploads/sites/47/2018/02/Displaced_Workers_WeRobot.pdf.

Elish, m c. "Don't Call AI Magic." *Data & Society: Points* (blog), January 17, 2018. https://points.datasociety.net/dont-call-ai-magic-142da16db408.

Elish, M. C., and danah boyd. "Situating Methods in the Magic of Big Data and AI." *Communication Monographs* 85, no. 1 (January 2, 2018): 57–80. https://doi.org/10.1080/03637751.2017.1375130.

Elish, Madeleine Clare. "Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction (Pre-Print)." *Engaging Science, Technology, and Society (Pre-Print)*, March 1, 2019. https://doi.org/10.2139/ssrn.2757236.

———. "The Stakes of Uncertainty: Developing and Integrating Machine Learning in Clinical Care." *EPIC* (blog), January 13, 2019. https://www.epicpeople.org/machine-learning-clinical-care/.

"Ethical Approaches to Closed Messaging Research: Considerations in Democratic Contexts – Election Standards." Accessed March 4, 2022. https://electionstandards.cartercenter.org/verifying-elections-misinfocon2020/ethical-approaches-to-closed-messaging-research-considerations-in-democratic-contexts/.

Fazelpour, Sina, and Zachary C. Lipton. "Algorithmic Fairness from a Non-Ideal Perspective." *ArXiv:2001.09773 [Cs, Stat]*, January 8, 2020. http://arxiv.org/abs/2001.09773.

Fjeld, Jessica, Nele Achten, Hannah Hilligoss, Adam Nagy, and Madhulika Srikumar. "Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI." SSRN Scholarly Paper. Rochester, NY: Social Science Research Network, January 15, 2020. https://doi.org/10.2139/ssrn.3518482.

François, Camille, Ben Nimmo, and C. Shawn Eib. "The IRACopyPasta Campaign." Graphika, October 2019. https://public-assets.graphika.com/reports/graphika_report_copypasta.pdf.

Frank, Morgan R., David Autor, James E. Bessen, Erik Brynjolfsson, Manuel Cebrian, David J. Deming, Maryann Feldman, et al. "Toward Understanding the Impact of Artificial Intelligence on Labor." *Proceedings of the National Academy of Sciences* 116, no. 14 (April 2, 2019): 6531–39. https://doi.org/10.1073/pnas.1900949116.

Frank, Morgan R., Lijun Sun, Manuel Cebrian, Hyejin Youn, and Iyad Rahwan. "Small Cities Face Greater Impact from Automation." *Journal of The Royal Society Interface* 15, no. 139 (February 28, 2018): 20170946. https://doi.org/10.1098/rsif.2017.0946.

Franz, Peter J., Erik C. Nook, Patrick Mair, and Matthew K. Nock. "Using Topic Modeling to Detect and Describe Self-Injurious and Related Content on a Large-Scale Digital Platform." *Suicide & Life-Threatening Behavior* 50, no. 1 (February 2020): 5–18. https://doi.org/10.1111/sltb.12569.

Friedberg, Brian, and Joan Donovan. "On the Internet, Nobody Knows You're a Bot: Pseudoanonymous Influence Operations and Networked Social Movements." *Journal of Design and Science*, no. 6 (August 7, 2019). https://doi.org/10.21428/7808da6b.45957184.

Friedberg, Joan Donovan, Brian. “Opinion: With This Statement, I Give Notice That Instagram Owns Your Soul.” *BuzzFeed News* (blog), August 23, 2019. https://www.buzzfeednews.com/article/joandonovan/i-give-notice-that-instagram-owns-you.

Friedler, Sorelle A., Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P. Hamilton, and Derek Roth. “A Comparative Study of Fairness-Enhancing Interventions in Machine Learning.” *ArXiv:1802.04422 [Cs, Stat]*, February 12, 2018. http://arxiv.org/abs/1802.04422.

Garg, Saurabh, Yifan Wu, Sivaraman Balakrishnan, and Zachary C. Lipton. “A Unified View of Label Shift Estimation.” In *34th Conference on Neural Information Processing Systems (NeurIPS 2020)*. Vancouver, Canada, 2020. https://proceedings.neurips.cc/paper/2020/file/219e052492f4008818b8adb6366c7ed6-Paper.pdf.

Gasser, Urs, and Virgilio A.F. Almeida. “A Layered Model for AI Governance.” *IEEE Internet Computing* 21, no. 6 (November 2017): 58–62. https://doi.org/10.1109/MIC.2017.4180835.

Goussac, Netta. “Safety Net or Tangled Web: Legal Reviews of AI in Weapons and War-Fighting.” *Humanitarian Law & Policy Blog* (blog), April 18, 2019. https://blogs.icrc.org/law-and-policy/2019/04/18/safety-net-tangled-web-legal-reviews-ai-weapons-war-fighting/.

Hasse, Alexa, Sandra Cortesi, Andres Lombana Bermudez, and Urs Gasser. “Youth and Artificial Intelligence: Where We Stand.” *Spotlight Series*, May 2019. https://dash.harvard.edu/handle/1/40268058.

Hellmann, Melissa. “A Tale of Two AI Cities: The Seattle Connection to Israel’s Surveillance Network.” *The Seattle Times*, April 18, 2020, sec. Technology. https://www.seattletimes.com/business/technology/washington-state-tech-giants-share-strong-links-with-israeli-ai-startups-fueling-global-surveillance-networks-but-also-raising-ethical-concerns/.

———. “AI 101: What Is Artificial Intelligence and Where Is It Going?” *The Seattle Times*, September 21, 2019, sec. Technology. https://www.seattletimes.com/business/technology/ai-101-what-is-artificial-intelligence-and-where-is-it-going/.

———. “AI Event in Seattle Brings Together Japanese Companies and US. Innovators.” *The Seattle Times*, July 27, 2019, sec. Technology. https://www.seattletimes.com/business/technology/ai-event-brings-together-japanese-companies-and-us-innovators/.

———. “AI Is Here to Stay, but Are We Sacrificing Safety and Privacy? A Free Public Seattle U Course Will Explore That.” *The Seattle Times*, February 11, 2020, sec. Technology. https://www.seattletimes.com/business/technology/seattle-university-launches-public-course-to-bring-the-ethics-of-artificial-intelligence-to-the-masses/.

———. “AI Software Beats Top Poker Players in Milestone of Computing.” *The Seattle Times*, July 11, 2019, sec. Technology. https://www.seattletimes.com/business/technology/ai-software-beats-top-poker-players-in-milestone-of-computing/#:~:text=But%20now%20researchers%20at%20Facebook,em%20for%20the%20first%20time.

———. “Amazon Conference Showcases Robots and Social Uses of Artificial Intelligence.” *The Seattle Times*, June 7, 2019, sec. Technology. https://www.seattletimes.com/business/amazon/amazon-conference-showcases-robots-and-social-uses-of-artificial-intelligence/.

———. “Amazon Speaks out in Favor of US. Regulating Facial-Recognition Technology.” *The Seattle Times*, June 11, 2019, sec. Technology. https://www.seattletimes.com/business/amazon/amazon-speaks-out-in-favor-of-regulating-facial-recognition/.

———. “Amazon Workers Bring Parents to Work.” *The Seattle Times*, September 13, 2019, sec. Technology. https://www.seattletimes.com/business/technology/amazon-workers-bring-parents-to-work/.

———. “Artist Works to Merge Artificial Intelligence and Art.” *The Seattle Times*, August 2, 2019, sec. Technology. https://www.seattletimes.com/business/technology/artist-works-to-merge-artificial-intelligence-and-art/.

———. “As Seattle’s New Hotels Roll out Automation to Serve Guests, Workers Worry.” *The Seattle Times*, May 18, 2019, sec. Technology. https://www.seattletimes.com/business/technology/as-seattles-new-hotels-roll-out-automation-to-serve-guests-workers-worry/.

———. “Augmented Writing Technology: A Writer’s Friend or Foe?” *The Seattle Times*, April 30, 2019, sec. Technology. https://www.seattletimes.com/business/technology/augmented-writing-technology-a-writers-friend-or-foe/.

———. “Axon Won’t Link Police Body Cameras to Facial-Recognition Technology.” *The Seattle Times*, June 27, 2019, sec. Technology. https://www.seattletimes.com/business/technology/axon-wont-link-police-body-cameras-to-facial-recognition-technology/.

———. “Bellevue Startup Uses Artificial Intelligence to Help People Learning a New Language.” *The Seattle Times*, October 29, 2019, sec. Technology. https://www.seattletimes.com/business/technology/bellevue-startup-uses-artificial-intelligence-to-help-english-learners-pronunciation/.

———. “Bonney Lake Walmart’s Robot Zips along in Tech Revolution That’s Raising Big Questions for Workers.” *The Seattle Times*, May 6, 2019, sec. Technology. https://www.seattletimes.com/business/technology/walmarts-push-for-advanced-technology-has-workers-asking-how-theyll-coexist-with-robots/.

———. “Conservationists Harness AI to Help Wolverine Recovery in Washington.” *The Seattle Times*, September 22, 2019, sec. Technology. https://www.seattletimes.com/business/technology/machine-learning-methods-harness-ai-to-help-wolverine-recovery/.

———. “‘Crowdworking’ Provides the Humans Who Train Artificial Intelligence.” *The Seattle Times*, July 26, 2019, sec. Technology. https://www.seattletimes.com/business/technology/crowdworking-provides-the-humans-who-train-artificial-intelligence-often-for-minimal-pay/.

———. “Group Calls for Federal Ban on Government Use of Facial-Recognition Technology Because It Misidentifies People of Color, Women.” *The Seattle Times*, July 9, 2019, sec. Technology. https://www.seattletimes.com/business/technology/advocacy-group-launches-national-campaign-to-ban-facial-recognition-technology-from-government-use/.

———. “IBM’s Watson Center Pitches AI for Everyone, from Chefs to Engineers | The A.I. Age.” *The Seattle Times*, January 26, 2020, sec. Technology. https://www.seattletimes.com/business/technology/ibms-watson-center-pitches-ai-for-everyone-from-chefs-to-engineers/.

———. “In Mount Vernon, Paccar Develops Automated Trucks That Could Revolutionize Long-Haul Industry.” *The Seattle Times*, April 12, 2020, sec. Technology. https://www.seattletimes.com/business/technology/automated-trucks-could-be-the-future-of-long-haul-trucking-but-cast-an-uncertain-future-for-truckers/.

———. “Inventions Including AI-Assisted Cat Door Highlighted at GeekWire Summit.” *The Seattle Times*, October 9, 2019, sec. Technology. https://www.seattletimes.com/business/technology/inventions-including-ai-assisted-cat-door-highlighted-at-geekwire-summit/.

———. “Magnolia Residents’ AI-Powered Surveillance Camera Tracks People, Cars at Entrance to Neighborhood, Experts Caution Bias.” *The Seattle Times*, December 7, 2019, sec. Technology. https://www.seattletimes.com/business/technology/magnolia-residents-ai-powered-surveillance-camera-tracks-people-cars-at-entrance-to-neighborhood-experts-caution-bias/.

———. “Microsoft Announces New AI for Accessibility Grantees.” *The Seattle Times*, May 15, 2019, sec. Technology. https://www.seattletimes.com/business/technology/microsoft-announces-new-ai-for-accessibility-grantees/.

———. “Q&A: Microsoft’s Lili Cheng Talks about Emotionally Intelligent Machines.” *The Seattle Times*, August 4, 2019, sec. Technology. https://www.seattletimes.com/business/technology/qa-microsofts-lili-cheng-talks-about-emotionally-intelligent-machines/.

———. “Seattle AI Lab’s Free Search Engine Aims to Accelerate Scientific Breakthroughs.” *The Seattle Times*, October 23, 2019, sec. Technology. https://www.seattletimes.com/business/technology/seattle-ai-labs-free-search-engine-aims-to-accelerate-scientific-breakthroughs/#:~:text=Seattle%20AI%20lab's%20free%20search%20engine%20aims%20to%20accelerate%20scientific%20breakthroughs,-Oct.&text=The%20Seattle%2Dbased%20Allen%20Institute,social%2C%20interdisciplinary%20and%20social%20sciences.

———. “Seattle Company Is Using Artificial Intelligence to Make Pizza; Check out the Assembly Line.” *The Seattle Times*, October 1, 2019, sec. Technology. https://www.seattletimes.com/business/technology/seattle-based-food-tech-company-finds-a-new-frontier-for-artificial-intelligence-pizza-production/.

———. “Seattle Faith Groups Reckon with AI — and What It Means to Be ‘Truly Human.’” *The Seattle Times*, November 10, 2019, sec. Technology. https://www.seattletimes.com/business/technology/seattle-faith-groups-reckon-with-ai-and-what-it-means-to-be-truly-human/.

———. “Seattle Startup 98point6 Puts Medical AI to Work with Sam’s Club.” *The Seattle Times*, September 26, 2019, sec. Technology. https://www.seattletimes.com/business/technology/seattle-startup-98point6-puts-medical-ai-to-work-with-sams-club/.

———. “Seattle’s Oversight of Surveillance Technology Is Moving Forward Slowly.” *The Seattle Times*, June 4, 2019, sec. Technology. https://www.seattletimes.com/business/technology/seattles-oversight-of-surveillance-technology-is-moving-forward-slowly/.

———. “Special Sunglasses, License-Plate Dresses: How to Be Anonymous in the Age of Surveillance.” *The Seattle Times*, January 12, 2020, sec. Technology. https://www.seattletimes.com/business/technology/special-sunglasses-license-plate-dresses-juggalo-face-paint-how-to-be-anonymous-in-the-age-of-surveillance/.

———. “Tech and Police Groups Urge Lawmakers to Not Ban Facial-Recognition Technology.” *The Seattle Times*, September 27, 2019, sec. Technology. https://www.seattletimes.com/business/technology/tech-and-police-groups-urge-lawmakers-not-to-ban-facial-recognition/.

———. “Tech Leaders and policymakers Talk Regulation at GeekWire Summit.” *The Seattle Times*, October 8, 2019, sec. Technology. https://www.seattletimes.com/business/technology/tech-leaders-and-policy-makers-talk-regulation-at-geekwire-summit/.

———. “Video Recap: Experts Discuss AI and the Future of Work.” *The Seattle Times*, September 23, 2019, sec. Technology. https://www.seattletimes.com/business/technology/join-us-for-a-conversation-on-ai-and-the-future-of-work/.

———. “When Convenience Meets Surveillance: AI at the Corner Store.” *The Seattle Times*, June 30, 2019, sec. Technology. https://www.seattletimes.com/business/technology/when-convenience-meets-surveillance-ai-at-the-corner-store/.

———. “White Collar Workers Will Be Most Affected by AI in the New Economy, Study Suggests.” *The Seattle Times*, November 19, 2019, sec. Technology. https://www.seattletimes.com/business/technology/study-suggests-white-collar-workers-will-be-most-affected-in-the-new-economy/#:~:text=Webb's%20analysis%20revealed%20that%20740,by%20AI%2C%20the%20report%20noted.

———. “Will Artificial Intelligence Make Work Better — or Worse? Seattle Times Event Explores the Future of Work.” *The Seattle Times*, November 11, 2019, sec. Technology. https://www.seattletimes.com/business/technology/will-artificial-intelligence-make-work-better-or-worse-seattle-times-event-explores-the-future-of-work/.

———. “With AI and Other Tech, Anat Caspi Focuses on Helping People with Disabilities.” *The Seattle Times*, August 4, 2019, sec. Technology. https://www.seattletimes.com/business/technology/anat-caspi-applies-artificial-intelligence-to-technology-focused-on-people-with-disabilities/.

Hessekiel, Kira H., Eliot Kim, James E. Tierney, Jonathan Y. Yang, and Christopher Bavitz. “AGTech Forum Briefing Book: State Attorneys General and Artificial Intelligence.” Berkman Klein Center for Internet & Society at Harvard University, 2018. https://dash.harvard.edu/handle/1/37184705.

Hickert, Cameron, and Jeffrey Ding. “Read What Top Chinese Officials Are Hearing About AI Competition and Policy.” *New America* (blog), November 29, 2018. http://newamerica.org/cybersecurity-initiative/digichina/blog/read-what-top-chinese-officials-are-hearing-about-ai-competition-and-policy/.

Ho, Scarlett. “China’s AI Efforts Suggest Tactics in New ‘Self-Reliance’ Push.” *DigiChina* (blog). Accessed March 2, 2022. https://digichina.stanford.edu/work/chinas-ai-efforts-suggest-tactics-in-new-self-reliance-push/.

Holland, Sarah, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. “The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards.” *ArXiv:1805.03677 [Cs]*, May 9, 2018. http://arxiv.org/abs/1805.03677.

Horsley, Jamie P. “How Will China’s Privacy Law Apply to the Chinese State?” *DigiChina* (blog). Accessed March 2, 2022. https://digichina.stanford.edu/work/how-will-chinas-privacy-law-apply-to-the-chinese-state/.

Huang, Yehan, and Mingli Shi. “Top Scholar Zhou Hanhua Illuminates 15+ Years of History Behind China’s Personal Information Protection Law.” *DigiChina* (blog). Accessed March 2, 2022. https://digichina.stanford.edu/work/top-scholar-zhou-hanhua-illuminates-15-years-of-history-behind-chinas-personal-information-protection-law/.

Ito, Joi. “What the Boston School Bus Schedule Can Teach Us About AI.” *Wired*. Accessed March 1, 2022. https://www.wired.com/story/joi-ito-ai-and-bus-routes/.

Jaroszewski, Adam C., Robert R. Morris, and Matthew K. Nock. “Randomized Controlled Trial of an Online Machine Learning-Driven Risk Assessment and Intervention Platform for Increasing the Use of Crisis Services.” *Journal of Consulting and Clinical Psychology* 87, no. 4 (April 2019): 370–79. https://doi.org/10.1037/ccp0000389.

Jayaram, Malavika. “Findings from Our AI Workshops.” *Digital Asia Hub* (blog), 2017. https://www.digitalasiahub.org/portfolio/findings-from-our-ai-workshops/.

Just, Marcel Adam, Lisa Pan, Vladimir L. Cherkassky, Dana L. McMakin, Christine Cha, Matthew K. Nock, and David Brent. “Machine Learning of Neural Representations of Suicide and Emotion Concepts Identifies Suicidal Youth.” *Nature Human Behaviour* 1 (2017): 911–19. https://doi.org/10.1038/s41562-017-0234-y.

Kaushik, Divyansh, Eduard Hovy, and Zachary Lipton. “Learning The Difference That Makes A Difference With Counterfactually-Augmented Data,” 2019. https://openreview.net/forum?id=Sklgs0NFvr.

Kessler, Ronald C., Samantha L. Bernecker, Robert M. Bossarte, Alex R. Luedtke, John F. McCarthy, Matthew K. Nock, Wilfred R. Pigeon, et al. “The Role of Big Data Analytics in Predicting Suicide.” In *Personalized Psychiatry*, 77–98, 2019. https://books.google.hu/books?id=zMOHDwAAQBAJ&pg=PA77&lpg=PA77&dq=Kessler,+R.C.,+Bernecker,+S.L.,+Bossarte,+R.M.,+Luedtke,+A.R.,+McCarthy,+J.F.,+Nock,+M.K.,+Pigeon,+W.+R.,+Petukhova,+M.V.,+Sadikova,+E.,+VanderWeele,+T.J.,+Zuromski,+K.L.,+%26+Zaslavsky,+A.M.+(2019).+The+role+of+big+data+analytics+in+predicting+suicide.+Personalized+Psychiatry,+77-98.&source=bl&ots=B6Pcm_NoJ4&sig=ACfU3U1sWSnVFJwtaqrh1TpIUI8X-Y_ajQ&hl=en&sa=X&ved=2ahUKEwjhlbjCkKX2AhVE_bsIHT05DrgQ6AF6BAgCEAM#v=onepage&q=Kessler%2C%20R.C.%2C%20Bernecker%2C%20S.L.%2C%20Bossarte%2C%20R.M.%2C%20Luedtke%2C%20A.R.%2C%20McCarthy%2C%20J.F.%2C%20Nock%2C%20M.K.%2C%20Pigeon%2C%20W.%20R.%2C%20Petukhova%2C%20M.V.%2C%20Sadikova%2C%20E.%2C%20VanderWeele%2C%20T.J.%2C%20Zuromski%2C%20K.L.%2C%20%26%20Zaslavsky%2C%20A.M.%20(2019).%20The%20role%20of%20big%20data%20analytics%20in%20predicting%20suicide.%20Personalized%20Psychiatry%2C%2077-98.&f=false.

Kleiber, Shannon Henry. “Should We Flee Social Media Or Fix It?” To The Best Of Our Knowledge, August 26, 2019. https://www.ttbook.org/interview/should-we-flee-social-media-or-fix-it.

Klein, Max, Julia Kamin, and J. Nathan Matias. “6 Ideas to Strengthen Wikipedia(s) with Citizen Behavioral Science.” *Citizens and Technology Lab* (blog), November 22, 2019. https://citizensandtech.org/2019/11/research-summit-with-wikimedians/.

Larsen, Benjamin. “Drafting China’s National AI Team for Governance.” *DigiChina* (blog). Accessed March 2, 2022. https://digichina.stanford.edu/work/drafting-chinas-national-ai-team-for-governance/.

Laux, Johann Moritz. *Public Epistemic Authority*. 1st ed. Grundlagen Der Rechtswissenschaft. Tübingen: Mohr Siebeck, 2021.

Laux, Johann, Sandra Wachter, and Brent Mittelstadt. “Neutralizing Online Behavioural Advertising: Algorithmic Targeting with Market Power as an Unfair Commercial Practice.” *Common Market Law Review* 58, no. 3 (April 9, 2021). https://papers.ssrn.com/abstract=3822962.

Lee, Alexa. “Personal Data, Global Effects: China’s Draft Privacy Law in the International Context.” *DigiChina* (blog). Accessed March 2, 2022. https://digichina.stanford.edu/work/personal-data-global-effects-chinas-draft-privacy-law-in-the-international-context/.

———. “The Future of Taiwan in US.-China Technology Competition.” *DigiChina* (blog). Accessed March 2, 2022. https://digichina.stanford.edu/work/the-future-of-taiwan-in-u-s-china-technology-competition-2/.

Lee, Alexa, Samm Sacks, Rogier Creemers, Mingli Shi, and Graham Webster. “China’s Draft Privacy Law Adds Platform Self-Governance, Solidifies CAC’s Role.” *DigiChina* (blog). Accessed March 2, 2022. https://digichina.stanford.edu/work/chinas-draft-privacy-law-adds-platform-self-governance-solidifies-cacs-role/.

Lee, Alexa, Mingli Shi, Qiheng Chen, Jamie P. Horsley, Kendra Schaefer, Rogier Creemers, and Graham Webster. “Seven Major Changes in China’s Finalized Personal Information Protection Law.” *DigiChina* (blog). Accessed March 2, 2022. https://digichina.stanford.edu/work/seven-major-changes-in-chinas-finalized-personal-information-protection-law/.

Lehmann, Thomas. “AI Politics Is Local.” *DigiChina* (blog). Accessed March 2, 2022. https://digichina.stanford.edu/work/ai-politics-is-local/.

Lewis, D. “International Legal Regulation of the Employment of Artificial-Intelligence-Related Technologies in Armed Conflict.” *Moscow Journal of International Law*, no. 2 (November 19, 2020): 53–64. https://doi.org/10.24833/0869-0049-2020-2-53-64.

Lewis, Dustin A. “A Key Set of IHL Questions Concerning A.I.-Supported Decision-Making.” In *50 Proceedings of the Bruges Colloquium*, forthcoming.

———. “AI and Machine Learning Symposium: Why Detention, Humanitarian Services, Maritime Systems, and Legal Advice Merit Greater Attention.” *Opinio Juris* (blog), April 28, 2020. http://opiniojuris.org/2020/04/28/ai-and-machine-learning-symposium-ai-in-armed-conflict-why-detention-humanitarian-services-maritime-systems-and-legal-advice-merit-greater-attention/.

———. “An Enduring Impasse on Autonomous Weapons.” *Just Security* (blog), September 28, 2020. https://www.justsecurity.org/72610/an-enduring-impasse-on-autonomous-weapons/.

———. “Legal Reviews of Weapons, Means and Methods of Warfare Involving Artificial Intelligence: 16 Elements to Consider.” *Humanitarian Law & Policy Blog* (blog), March 21, 2019. https://blogs.icrc.org/law-and-policy/2019/03/21/legal-reviews-weapons-means-methods-warfare-artificial-intelligence-16-elements-consider/.

———. “On ‘Responsible A.I.’ in War: Exploring Precoditions for Respecting International Law in Armed Conflict.” In *Responsibile A.I.* Freiburg: Freiburg Institute for Advanced Studies, forthcoming.

———. “Preconditions.” In *Autonomous Cyber Capabilities under International Law*, edited by Rain Liivoja and Ann Väljataga. NATO Cooperative Cyber Defence Centre of Excellence, 2021. https://ccdcoe.org/uploads/2021/05/Autonomous-Cyber-Capabilities-under-International-Law.pdf.

———. “Three Pathways to Secure Greater Respect for International Law Concerning War Algorithms.” Cambridge, MA: Harvard Law School Program on International Law and Armed Conflict, 2020. https://pilac.law.harvard.edu/three-pathways-to-secure-greater-respect-for-international-law-concerning-war-algorithms.

Li, Qiang, and Dan Xie. “Legal Regulation of AI Weapons under International Humanitarian Law: A Chinese Perspective.” *Humanitarian Law & Policy Blog* (blog), May 2, 2019. https://blogs.icrc.org/law-and-policy/2019/05/02/ai-weapon-ihl-legal-regulation-chinese-perspective/.

Lipton, Zachary C., Alexandra Chouldechova, and Julian McAuley. “Does Mitigating ML’s Impact Disparity Require Treatment Disparity?” *ArXiv:1711.07076 [Cs, Stat]*, January 11, 2019. http://arxiv.org/abs/1711.07076.

Lucero, Karman. “In China, Planning Towards AI Policy Paralysis.” *DigiChina* (blog). Accessed March 2, 2022. https://digichina.stanford.edu/work/in-china-planning-towards-ai-policy-paralysis/.

Lum, Kristian, Chesa Boudin, and Megan Price. “The Impact of Overbooking on a Pre-Trial Risk Assessment Tool.” *ArXiv:2001.08793 [Stat]*, January 23, 2020. http://arxiv.org/abs/2001.08793.

Lum, Kristian, David B. Dunson, and James Johndrow. “Closer than They Appear: A Bayesian Perspective on Individual-Level Heterogeneity in Risk Assessment.” *ArXiv:2102.01135 [Stat]*, February 1, 2021. http://arxiv.org/abs/2102.01135.

Lum, Kristian, and Tarak Shah. “MEASURES OF FAIRNESS FOR NEW YORK CITY’S SUPERVISED RELEASE RISK ASSESSMENT TOOL.” Human Rights Data Analysis Group, n.d. https://hrdag.org/wp-content/uploads/2019/09/2019-HRDAG-measures-of-fairness-CJA-1.pdf.

Luo, Yan, Samm Sacks, Abigail Coplin, and Naomi Wilson. “Mapping US.–China Technology Decoupling.” *DigiChina* (blog). Accessed March 2, 2022. https://digichina.stanford.edu/work/mapping-u-s-china-technology-decoupling/.

Mateescu, Alexandra, and Madeleine Clare Elish. “AI in Context: The Labor of Integrating New Technologies.” New York: Data & Society Research Institute, January 30, 2019. https://datasociety.net/wp-content/uploads/2019/01/DataandSociety_AIinContext.pdf.

Matias, J. Nathan. “New Directions for Citizen Research and Action on Digital Power.” *Citizens and Technology Lab* (blog), December 12, 2019. https://citizensandtech.org/2019/12/new-directions-for-citizen-research-action/.

———. “Why We Need Industry-Independent Research on Tech & Society.” *Citizens and Technology Lab* (blog), January 7, 2020. https://citizensandtech.org/2020/01/industry-independent-research/.

Matias, J. Nathan, Florence Devouard, Julia Kamin, and Max Klein. “Study Results: Changing How Wikipedia Represents Africa With Photo Recruitment Campaigns.” *Citizens and Technology Lab* (blog), August 6, 2020. https://citizensandtech.org/2020/08/study-results-wikilovesafrica-2020/.

Matias, J. Nathan, Taylor Simko, and Marianne Reddan. “Study Results: Reducing the Silencing Role of Harassment in Online Feminism Discussions.” *Citizens and Technology Lab* (blog), June 25, 2020. https://citizensandtech.org/2020/06/reducing-harassment-impacts-in-feminism-online/.

McGregor, Lorna. “The Need for Clear Governance Frameworks on Predictive Algorithms in Military Settings.” *Humanitarian Law & Policy Blog* (blog), March 28, 2019. https://blogs.icrc.org/law-and-policy/2019/03/28/need-clear-governance-frameworks-predictive-algorithms-military-settings/.

McInnis, Katie. “Data from the People, for the People: Consumer Reports and CAT Lab’s Distributed Research….” *Digital Lab at Consumer Reports* (blog), November 26, 2019. https://medium.com/cr-digital-lab/data-from-the-people-for-the-people-consumer-reports-and-cat-labs-distributed-research-8e69c39b1d69.

McKeown, Alex, Andrew Turner, Zuzanna Angehrn, Dianne Gove, Amanda Ly, Clementine Nordon, Mia Nelson, et al. “Health Outcome Prioritization in Alzheimer’s Disease: Understanding the Ethical Landscape.” *Journal of Alzheimer’s Disease: JAD* 77, no. 1 (2020): 339–53. https://doi.org/10.3233/JAD-191300.

Milano, Silvia, Brent Mittelstadt, Sandra Wachter, and Christopher Russell. “Epistemic Fragmentation Poses a Threat to the Governance of Online Targeting.” *Nature Machine Intelligence* 3, no. 6 (June 2021): 466–72. https://doi.org/10.1038/s42256-021-00358-3.

Milano, Silvia, Mariarosaria Taddeo, and Luciano Floridi. “Ethical Aspects of Multi-Stakeholder Recommendation Systems.” *The Information Society* 37, no. 1 (January 1, 2021): 35–45. https://doi.org/10.1080/01972243.2020.1832636.

Miller, John, Smitha Milli, and Moritz Hardt. “Strategic Classification Is Causal Modeling in Disguise.” In *Proceedings of the 37th International Conference on Machine Learning*, 6917–26. PMLR, 2020. https://proceedings.mlr.press/v119/miller20b.html.

Milli Network. *IAW2020 Milli Sessions: Day 5, Jun 12: Archives and Crises*, 2020. https://www.youtube.com/watch?v=mtIaD-AvG_U.

Mitchell, Shira, Eric Potash, Solon Barocas, Alexander D’Amour, and Kristian Lum. “Prediction-Based Decisions and Fairness: A Catalogue of Choices, Assumptions, and Definitions.” *Annual Review of Statistics and Its Application* 8, no. 1 (March 7, 2021): 141–63. https://doi.org/10.1146/annurev-statistics-042720-125902.

Nelson, Mei. “Debating China’s AI Path: ‘Alternative Routes,’ or ‘Overtaking on the Curve’?” *DigiChina* (blog). Accessed March 2, 2022. https://digichina.stanford.edu/work/debating-chinas-ai-path-alternative-routes-or-overtaking-on-the-curve/.

Nimmo, Ben. “UK Trade Leaks.” Graphika, December 2019. https://public-assets.graphika.com/reports/graphika_report_uk_trade_leaks_updated_12.12.pdf.

Nimmo, Ben, C. Shawn Eib, and L. Tamora. “Cross-Platform Span Network Targeted Hong Kong Protests.” Graphika, September 2019. https://public-assets.graphika.com/reports/graphika_report_spamouflage.pdf.

Nimmo, Ben, C. Shawn Eib, L. Tamora, Kate Johnson, Ian Smith, Eto Buziashvili, Alyssa Kann, Kanishk Karan, Esteban Ponce de León Rosas, and Max Rizzuto. “#OperationFFS: Fake Face Swarm.” Graphika, December 2019. https://public-assets.graphika.com/reports/graphika_report_operation_ffs_fake_face_storm.pdf.

Nimmo, Ben, Camille François, C. Shawn Eib, and L. Tamora. “From Russia with Blogs: GRU Operators Leveraged Blogs, Social Media Accounts and Private Messaging to Reach Audiences Across Europe.” February 2020: Graphika, n.d. https://public-assets.graphika.com/reports/graphika_report_from_russia_with_blogs.pdf.

OECD. “Public Policy Considerations.” In *Artificial Intelligence in Society*. Paris: OECD Publishing, 2019. https://doi.org/10.1787/eedfee77-en.

Panofsky, Aaron, and Joan Donovan. “Genetic Ancestry Testing among White Nationalists: From Identity Repair to Citizen Science.” *Social Studies of Science* 49, no. 5 (October 1, 2019): 653–81. https://doi.org/10.1177/0306312719861434.

Paris, Britt, and Joan Donovan. “Deepfakes and Cheap Fakes: The Manipulation of Audio and Visual Evidence.” Data & Society Research Institute, September 18, 2019. https://datasociety.net/wp-content/uploads/2019/09/DS_Deepfakes_Cheap_FakesFinal-1-1.pdf.

Parkin, Siodhbhra. “How AI Can Better Serve People With Disabilities in China.” *DigiChina* (blog). Accessed March 2, 2022. https://digichina.stanford.edu/work/how-ai-can-better-serve-people-with-disabilities-in-china/.

Parra, Paulette Desormeaux. “La polémica tras el algoritmo que busca mejorar la equidad en el acceso a la educación en Chile.” *Chequeado* (blog), September 30, 2020. https://chequeado.com/investigaciones/la-polemica-tras-el-algoritmo-que-busca-mejorar-la-equidad-en-el-acceso-a-la-educacion-en-chile/.

Pennycook, Gordon, Ziv Epstein, Mohsen Mosleh, Antonio A. Arechar, Dean Eckles, and David G. Rand. “Shifting Attention to Accuracy Can Reduce Misinformation Online.” *Nature* 592, no. 7855 (April 2021): 590–95. https://doi.org/10.1038/s41586-021-03344-2.

Pennycook, Gordon, and David Rand. “Opinion | Why Do People Fall for Fake News?” *The New York Times*, January 19, 2019, sec. Opinion. https://www.nytimes.com/2019/01/19/opinion/sunday/fake-news.html.

Pennycook, Gordon, and David G. Rand. “Fighting Misinformation on Social Media Using Crowdsourced Judgments of News Source Quality.” *PNAS* 116, no. 7 (January 28, 2019). https://doi.org/10.1073/pnas.1806781116.

Porter, Ethan, and Thomas J. Wood. “The Global Effectiveness of Fact-Checking: Evidence from Simultaneous Experiments in Argentina, Nigeria, South Africa, and the United Kingdom.” *PNAS* 118, no. 7 (September 10, 2021). https://doi.org/10.1073/pnas.2104235118.

Prabhakar, Tarunima, and Denny George. “Considerations in Archiving Misinformation from Encrypted Messaging Apps,” October 1, 2019. https://cyber.harvard.edu/sites/default/files/2019-12/Tattle_Disinformation_Workshop_revised.pdf.

Prabhakar, Tarunima, Anushree Gupta, Kruttika Nadig, and Denny George. “Check Mate: Prioritizing User Generated Multi-Media Content for Fact-Checking.” *Proceedings of the International AAAI Conference on Web and Social Media* 15 (May 22, 2021): 1025–33. https://ojs.aaai.org/index.php/ICWSM/article/view/18126.

Princeton University. “Automated Healthcare App Case Study: 1,” 2018. https://aiethics.princeton.edu/wp-content/uploads/sites/587/2018/10/Princeton-AI-Ethics-Case-Study-1.pdf.

———. “Dynamic Sound Identification Case Study: 2,” 2018. https://aiethics.princeton.edu/wp-content/uploads/sites/587/2018/10/Princeton-AI-Ethics-Case-Study-2.pdf.

———. “Hiring by Machine Case Study: 5,” 2018. https://aiethics.princeton.edu/wp-content/uploads/sites/587/2018/12/Princeton-AI-Ethics-Case-Study-5.pdf.

———. “Law Enforcement Chatbots Case Study: 4,” 2018. https://aiethics.princeton.edu/wp-content/uploads/sites/587/2018/10/Princeton-AI-Ethics-Case-Study-4.pdf.

———. “Optimizing Schools Case Study: 3,” 2018. https://aiethics.princeton.edu/wp-content/uploads/sites/587/2018/10/Princeton-AI-Ethics-Case-Study-3.pdf.

———. “Public Sector Data Analytics Case Study: 6,” 2018. https://aiethics.princeton.edu/wp-content/uploads/sites/587/2018/10/Princeton-AI-Ethics-Case-Study-6.pdf.

Radin, Sasha. “Expert Views on the Frontiers of Artificial Intelligence and Conflict.” *Humanitarian Law & Policy Blog* (blog), March 19, 2019. https://blogs.icrc.org/law-and-policy/2019/03/19/expert-views-frontiers-artificial-intelligence-conflict/.

Rahwan, Iyad. “Society-in-the-Loop: Programming the Algorithmic Social Contract.” *Ethics and Information Technology* 20, no. 1 (March 2018): 5–14. https://doi.org/10.1007/s10676-017-9430-8.

———. “Towards Scalable Governance: Sensemaking and Cooperation in the Age of Social Media.” *Philosophy & Technology* 30, no. 2 (2017). https://doi.org/10.1007/s13347-016-0246-y.

Rahwan, Iyad, Manuel Cebrian, Nick Obradovich, Josh Bongard, Jean-François Bonnefon, Cynthia Breazeal, Jacob W. Crandall, et al. “Machine Behaviour.” *Nature* 568, no. 7753 (April 2019): 477–86. https://doi.org/10.1038/s41586-019-1138-y.

Raso, Filippo A., Hannah Hilligoss, Vivek Krishnamurthy, Christopher Bavitz, and Levin Kim. “Artificial Intelligence & Human Rights: Opportunities & Risks.” SSRN Scholarly Paper. Rochester, NY: Social Science Research Network, September 25, 2018. https://doi.org/10.2139/ssrn.3259344.

Reisman, Dillon, Jason Schultz, Kate Crawford, and Meredith Whittaker. “Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability.” AI Now, April 2018. https://ainowinstitute.org/aiareport2018.pdf.

Russell, C., A. Elliott, and S. Law. “Object Localisation Using Perturbations on the Perceptual Ball.” In *Proceedings of WHI 2020, ICML Workshop*, 2020.

Russell, C., R. McGrath, and L. Costabello. “Learning Relevant Explanations.” In *Proceedings of WHI 2020, ICML Workshop*, 2020.

Sacks, Samm. “Samm Sacks Testifies Before House Foreign Affairs Committee on ‘Smart Competition’ With China.” *New America* (blog), May 10, 2019. http://newamerica.org/cybersecurity-initiative/digichina/blog/samm-sacks-testifies-house-foreign-affairs-committee-smart-competition-china/.

Sacks, Samm, Mingli Shi, Rogier Creemers, Cindy L, and Paul Triolo. “Public Security Ministry Aligns With Chinese Data Protection Regime in Draft Rules.” *New America* (blog), December 3, 2018. http://newamerica.org/cybersecurity-initiative/digichina/blog/public-security-ministry-aligns-chinese-data-protection-regime-draft-rules/.

Sacks, Samm, Mingli Shi, and Graham Webster. “The Evolution of China’s Data Governance Regime: A Timeline.” *New America* (blog), February 8, 2019. http://newamerica.org/cybersecurity-initiative/digichina/blog/china-data-governance-regime-timeline/.

Sacks, Samm, Graham Webster, and Qiheng Chen. “Five Important Takeaways From China’s Draft Data Security Law.” *DigiChina* (blog). Accessed March 2, 2022. https://digichina.stanford.edu/work/five-important-takeaways-from-chinas-draft-data-security-law/.

Schaefer, Kendra, Samm Sacks, and Xiaomeng Lu. “With Auto Data, China Buckles In for Security and Opens Up for Future Tech.” *DigiChina* (blog). Accessed March 2, 2022.

https://digichina.stanford.edu/work/with-auto-data-china-buckles-in-for-security-and-opens-up-for-future-tech/.

Schneider, Jordan. “Could an ‘AI Winter’ Be on the Horizon for China?” *DigiChina* (blog). Accessed March 2, 2022. https://digichina.stanford.edu/work/could-an-ai-winter-be-on-the-horizon-for-china/.

Shariff, Azim, Jean-François Bonnefon, and Iyad Rahwan. “Psychological Roadblocks to the Adoption of Self-Driving Vehicles.” *Nature Human Behaviour* 1, no. 10 (October 2017): 694–96. https://doi.org/10.1038/s41562-017-0202-6.

Shi, Mingli. “China’s Draft Privacy Law Both Builds On and Complicates Its Data Governance.” *DigiChina* (blog). Accessed March 2, 2022. https://digichina.stanford.edu/work/chinas-draft-privacy-law-both-builds-on-and-complicates-its-data-governance/.

———. “What China’s 2018 Internet Governance Tells Us About What’s Next.” *New America* (blog), January 28, 2019. http://newamerica.org/cybersecurity-initiative/digichina/blog/what-chinas-2018-internet-governance-tells-us-about-whats-next/.

Sohrawardi, Saniat Javid, Sovantharith Seng, and Akash Chintha. “DeFaking Deepfakes: Understanding Journalists’ Needs for Deepfake Detection,” n.d., 5.

Tai, Katharin. “Chinese Interagency Group Calls Out Apps for Illegally Collecting User Data.” *New America* (blog), July 29, 2019. http://newamerica.org/cybersecurity-initiative/digichina/blog/chinese-interagency-group-calls-out-apps-illegally-collecting-user-data/.

Tappin, Ben M., Gordon Pennycook, and David G. Rand. “Rethinking the Link between Cognitive Sophistication and Politically Motivated Reasoning.” *Journal of Experimental Psychology: General* 150, no. 6 (2021): 1095–1114. https://doi.org/10.1037/xge0000974.

Tarricone, Manuel. “¿Hasta qué punto pueden automatizarse las decisiones judiciales? Enterate cómo funciona el software que ya se usa en la Ciudad de Buenos Aires.” *Chequeado* (blog), September 30, 2020. https://chequeado.com/investigaciones/hasta-que-punto-pueden-automatizarse-las-decisiones-judiciales-enterate-como-funciona-el-software-que-ya-se-usa-en-la-ciudad-de-buenos-aires/.

Technology and Social Change project. “Media Manipulation Casebook.” Media Manipulation Casebook. Accessed March 1, 2022. https://mediamanipulation.org/homepage.

Toner, Helen, and Lorand Laskai. “Can China Grow Its Own AI Tech Base?” *DigiChina* (blog). Accessed March 2, 2022. https://digichina.stanford.edu/work/can-china-grow-its-own-ai-tech-base/.

Toner, Helen, Paul Triolo, and Rogier Creemers. “Experts Examine China’s Pioneering Draft Algorithm Regulations.” *DigiChina* (blog). Accessed March 2, 2022. https://digichina.stanford.edu/work/experts-examine-chinas-pioneering-draft-algorithm-regulations/.

Torous, John, Mark E. Larsen, Colin Depp, Theodore D. Cosco, Ian Barnett, Matthew K. Nock, and Joe Firth. “Smartphones, Sensors, and Machine Learning to Advance Real-Time Prediction and Interventions for Suicide Prevention: A Review of Current Progress and Next Steps.” *Current Psychiatry Reports* 20, no. 7 (July 2018): 51. https://doi.org/10.1007/s11920-018-0914-y.

Toso, Matteo, Neill D F Campbell, and Chris Russell. “Fixing Implicit Derivatives: Trust-Region Based Learning of Continuous Energy Functions (Abridged),” n.d. https://anucvml.github.io/ddn-cvprw2020/papers/Toso_et_al_cvprw2020.pdf.

Triolo, Paul. “China’s World Internet Conference Struggles to Live Up to Its Name.” *New America* (blog), November 6, 2018. http://newamerica.org/cybersecurity-initiative/digichina/blog/chinas-world-internet-conference-struggles-to-live-up-to-its-name/.

Triolo, Paul, Elsa Kania, Jacqueline Musiitwa, Maarten Van Horenbeeck, Justin Sherman, Rui Zhong, Jessica Cussins Newman, and Charlotte Stix. “Online Symposium: Chinese Thinking on AI Security in Comparative Context.” *New America* (blog), February 21, 2019. http://newamerica.org/cybersecurity-initiative/digichina/blog/online-symposium-chinese-thinking-ai-security-comparative-context/.

Triolo, Paul, Samm Sacks, Graham Webster, and Rogier Creemers. “After 5 Years, China’s Cybersecurity Rules for Critical Infrastructure Come Into Focus.” *DigiChina* (blog). Accessed March 2, 2022.

https://digichina.stanford.edu/work/after-5-years-chinas-cybersecurity-rules-for-critical-infrastructure-come-into-focus/.

Triolo, Paul, and Graham Webster. “China’s Efforts to Build the Semiconductors at AI’s Core.” New America, December 7, 2018. http://newamerica.org/cybersecurity-initiative/digichina/blog/chinas-efforts-to-build-the-semiconductors-at-ais-core/.

Venkatasubramanian, Suresh. “Structural Disconnects between Algorithmic Decision-Making and the Law.” *Humanitarian Law & Policy Blog* (blog), April 25, 2019. https://blogs.icrc.org/law-and-policy/2019/04/25/structural-disconnects-algorithmic-decision-making-law/.

Wachter, Sandra. “Affinity Profiling and Discrimination by Association in Online Behavioural Advertising.” *Berkeley Technology Law Journal* 35, no. 2 (March 25, 2020). https://dx.doi.org/10.2139/ssrn.3388639.

Wachter, Sandra, Brent Mittelstadt, and Chris Russell. “Why Fairness Cannot Be Automated: Bridging the Gap Between EU Non-Discrimination Law and AI.” *Computer Law & Security Review* 41, no. 2021 (March 3, 2020): 105567. https://doi.org/10.2139/ssrn.3547922.

Wall Street Journal. *How Radicalization Online Can (And Can’t) Be Stopped | WSJ*, 2019. https://www.youtube.com/watch?v=A6984NNJyWQ.

Webster, Graham. “What the Huawei Executive’s Arrest Could Mean for the US.-China Relations.” *New America* (blog), December 11, 2018. http://newamerica.org/cybersecurity-initiative/digichina/blog/what-the-huawei-executives-arrest-could-mean-for-the-us-china-relations/.

Webster, Graham, and Samm Sacks. “Five Big Questions Raised by China’s New Draft Cross-Border Data Rules.” *New America* (blog), June 13, 2019. http://newamerica.org/cybersecurity-initiative/digichina/blog/five-big-questions-raised-chinas-new-draft-cross-border-data-ruless/.

Webster, Graham, Samm Sacks, and Paul Triolo. “Three Chinese Digital Economy Policies at Stake in the US.–China Talks.” *New America* (blog), April 2, 2019. http://newamerica.org/cybersecurity-initiative/digichina/blog/three-chinese-digital-economy-policies-at-stake-in-the-uschina-talks/.

Xiao, Muyi. “Human Resources Both Drive and Limit China’s Push for Automation.” *DigiChina* (blog). Accessed March 2, 2022. https://digichina.stanford.edu/work/human-resources-both-drive-and-limit-chinas-push-for-automation/.

Yau, Herman, Chris Russell, and Simon Hadfield. “What Did You Think Would Happen? Explaining Agent Behaviour Through Intended Outcomes.” *ArXiv:2011.05064 [Cs, Stat]*, November 10, 2020. http://arxiv.org/abs/2011.05064.

Yeung, Karen, and Adrian Weller. “How Is ‘Transparency’ Understood By Legal Scholars And The Machine Learning Community?” In *Being Profiled: Cogitas Ergo Sum*, 36–41. Amsterdam University Press, 2018. https://doi.org/10.1515/9789048550180-007.

Zhang, Amy X., Martin Robbins, Ed Bice, Sandro Hawke, David Karger, An Xiao Mina, Aditya Ranganathan, et al. “A Structured Response to Misinformation: Defining and Annotating Credibility Indicators in News Articles.” In *Companion of the The Web Conference 2018 on The Web Conference 2018 - WWW ’18*, 603–12. Lyon, France: ACM Press, 2018. https://doi.org/10.1145/3184558.3188731.

Zhang, Baobao. “Opinion | It’s 2043. We Need a New American Dream for the A.I. Revolution.” *The New York Times*, August 12, 2019, sec. Opinion. https://www.nytimes.com/2019/08/12/opinion/ubi-automation-ai.html.

———. “Public Opinion Toward Artificial Intelligence.” OSF Preprints, October 7, 2021. https://doi.org/10.31219/osf.io/284sm.

Zhang, Baobao, Markus Anderljung, Lauren Kahn, Noemi Dreksler, Michael C. Horowitz, and Allan Dafoe. “Ethics and Governance of Artificial Intelligence: Evidence from a Survey of Machine Learning Researchers.” *Journal of Artificial Intelligence Research* 71 (August 2, 2021): 591–666. https://doi.org/10.1613/jair.1.12895.

Zhang, Baobao, and Allan Dafoe. “Artificial Intelligence: American Attitudes and Trends.” Center for the Governance of AI (GovAI), January 2019. https://governanceai.github.io/US-Public-Opinion-Report-Jan-2019/executive-summary.html.

Zimmermann, Annette, and Bendert Zevenbergen. “AI Ethics: Seven Traps.” *Freedom to Tinker* (blog), March 25, 2019. https://freedom-to-tinker.com/2019/03/25/ai-ethics-seven-traps/.

Zittrain, Jonathan. “How to Exercise the Power You Didn’t Ask For.” *Harvard Business Review* (blog), September 19, 2018. https://perma.cc/W233-C7Q6.

———. “Opinion | Mark Zuckerberg Can Still Fix This Mess.” *The New York Times*, April 7, 2018, sec. Opinion. https://www.nytimes.com/2018/04/07/opinion/sunday/zuckerberg-facebook-privacy-congress.html.

Zwetsloot, Remco, Baobao Zhang, Noemi Dreksler, Lauren Kahn, Markus Anderljung, Allan Dafoe, and Michael C. Horowitz. “Skilled and Mobile: Survey Evidence of AI Researchers’ Immigration Preferences.” In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 1050–59. New York, NY, USA: Association for Computing Machinery, 2021. https://doi.org/10.1145/3461702.346261