PARTNERSHIP ON AI

PARTNERSHIP ON AI 2020 —ANNUAL REPORT—

PARTNERSHIP ON AI

Our Vision

A future where Artificial Intelligence empowers humanity by contributing to a more just, equitable, and prosperous world.

Our Mission

Bringing diverse voices together across global sectors, disciplines, and demographics so developments in AI advance positive outcomes for people and society.

The Partnership on AI (PAI) is an independent, nonprofit 501(c)(3) organization. It was originally established by a coalition of representatives from technology companies, civil society organizations, and academic institutions, and supported originally by multi-year grants from Apple, Amazon, Facebook, Google/DeepMind, IBM and Microsoft.

From Our Acting Executive Director:

It is impossible to look back on the last year and not think about the innumerable changes, great and small, that have transformed the world. At the same time, this act of reflection makes it even clearer what has remained the same. Perhaps above all, what has persisted is the critical need for community in our lives. In the field of artificial intelligence, the Partnership on AI (PAI) has worked from its very founding to meet this need.

With a Partner community that spans sectors, disciplines, and borders, PAI has always been an extraordinary organization. As we enter the new year, what stands out are the successes in consensus-building, alignment, and collaboration facilitated by this diverse set of perspectives. Like so many organizations, 2020 forced us to shift how we connected with each other, learning alongside our Partners as we all adjusted to new ways of achieving our goals together. While this has been a tough journey, it has also been a deeply rewarding one, in many ways bringing our community closer together than ever.

In the past year, we have also seen **PAI's unique position put us at the center of important conversations related to technology and society** as they become some of the most pressing issues of our time. As AI reshapes economies, political and justice systems, and other domains around the globe, we are proud to support an active coalition of organizations meaningfully collaborating to promote responsible AI. The moment that we find ourselves living through, with all of its challenges and opportunities, makes it even clearer how important the Partnership's work and mission are to our collective future.

In this year's annual report, you will find the vision, mission, and values statements we solidified in 2020 with the help of our Partner community. You will also learn more specifically about **the impact of PAI and our Partner community**. This includes the investigation of challenges to diversity and inclusion in the field of AI, the convocation of a diverse group of experts to chart a path forward for transparency in machine learning, and a database launched to document failures in AI and help prevent repeating the mistakes of the past.

Additionally, you will hear from some of **the Partners who make this community so special**. As you turn through these pages, I hope you will enjoy all of these highlights from the last year.

As Acting Executive Director, I look forward to PAI continuing to be a nexus for so many organizations considering the future of AI development and deployment. This would not be possible without the contributions of Terah Lyons, who departs as Executive Director after guiding PAI from its first steps to its greatest strides. I would also like to recognize PAI's engaged Board of Directors for their vision and leadership. In the coming weeks, we'll be in touch about work underway to renew PAI's strategic plan in consultation with Partners.

We thank all of our Partners for their insight, their enthusiasm, and, above all, their unwavering commitment to a future where AI empowers not just some, but every one of us.

Sincerely,
Rebecca Finlay
PAI Board Member and Acting Executive Director

February 10, 2021

Today, artificial intelligence is changing everything from the way we work to the way we see the world itself. Used responsibly, **we believe AI can create a future that is more just, prosperous, and equitable for all**. Guiding this technology so that it best serves everyone, however, is a task too great for anyone to accomplish alone.

By bringing together civil society organizations, technology companies, academic institutions, and many more members from across the globe, **the Partnership on AI connects the people building these systems with those impacted by them or studying their effects**. Our Partners' diverse perspectives also inform our original research, contributing to the rigorous development of best practices for AI.

In 2020, this research primarily fell under five Issue Areas: ABOUT ML; AI, Labor, and the Economy; AI and Media Integrity; Fairness, Transparency, and Accountability; and Safety-Critical AI.

ABOUT ML:

ABOUT ML seeks to establish an industry-wide norm for the documentation of machine learning systems, one that supports the goals of transparency, responsibility, and accountability. Additionally, our Methods for Inclusion project is investigating the barriers to communication between AI developers and the diverse communities their work impacts.

IMPACT: Last March, **the National Security Council cited ABOUT ML in their recommendations to Congress**, holding it up as a model for responsible AI documentation that government agencies should emulate. And in April, PAI researchers contributed to **a multistakeholder report** on how to ensure claims about AI systems are verifiable.

AI, Labor, and the Economy:

AI, Labor, and the Economy's Shared Prosperity Initiative dares to imagine a world where innovation works to enhance humanity's industriousness and creativity — not just supplant it. Whether AI makes the poor poorer or all of us richer, is a choice for us to make. In addition, the Responsible Sourcing initiative examines labor conditions within AI development itself, specifically focussing on the professionals who clean and label training data or otherwise contribute human judgment to AI systems.

IMPACT: To chart a course where AI's economic beneifts don't enrich the few at the expense of the many, **23 notable thinkers from around the globe were brought together** virtually last fall, identifying major topics of study for this emerging discipline of Responsible AI.

AI and Media Integrity:

While AI has ushered in an unprecedented era of knowledge-sharing online, it has also led to new categories of harmful digital content and extended their potential reach. PAI's Media Integrity program area directly addresses these critical challenges to the quality of public discourse and information. This includes ongoing work on the detection and labeling of manipulated media as well as upcoming research identifying potential threats, testing interventions, and exploring responsible content-ranking principles.

IMPACT: Last March, this work resulted in **a report offering six specific recommendations drawn from the Deepfake Detection Challenge**. In June, PAI published **a set of 12 principles** designers should follow when labeling manipulated media online, the result of an ongoing collaboration with First Draft.

Fairness, Transparency, and Accountability:

The Fairness, Transparency, and Accountability Issue Area encompasses PAI's extensive research concerning the intersections of AI with equity and social justice. This includes new and continuing initiatives examining algorithmic fairness, the criminal justice system, and diversity and inclusion as they relate to both the application of AI systems and the AI community itself.

IMPACT: In 2020, our Fairness, Transparency, and Accountability work resulted in **an issue brief explaining why the algorithmic PATTERN tool must not dictate federal prisoner transfers during the COVID-19 pandemic**, **a paper** offering an alternative legal framework for mitigating algorithmic bias, and **a new fellowship** studying barriers to diversity and inclusion in AI.

Safety-Critical AI:

As our lives are increasingly saturated with artificial intelligence systems, the safety of these systems becomes a vital consideration. The Safety-Critical AI Issue Area seeks to establish norms and technical foundations that will support the safe development and deployment of AI.

IMPACT: At NeurIPS 2020, PAI co-hosted **a workshop addressing open questions concerning responsible oversight of novel AI research**. And in October, PAI announced **a new competitive benchmark** for training non-destructive agents in an AI learning environment.

Learn more about all of PAI's program's on our projects page.

OUR YEAR IN NUMBERS

Number of Partners: **98**

- **26** Academic Partners
- **52** Nonprofit Partners
- **17** Industry Partners
- **3** Media Partners

13 Countries our Partners call home.

Workshops hosted in 2020: **28**

46 Talks given, including presentations at ACM FAccT 2020 and IBM's annual Think conference.

PAI Research Fellows in 2020: **11**

11 Papers accepted for publication by journals and conferences like The University of Chicago Law Review Online and FAccT.

Equity & Inclusion

As an independent body, PAI exists for the benefit of people and society. **We empower diverse voices** to participate from ideation through implementation, striving for fairness, equity, and inclusion.

Conviction & Dependability

Committed to tackling the hard questions through meaningful dialogue, research, insights, and guidance, **we're determined to maintain courage in the face of adversity**, and are dedicated to facilitating effective processes that lead to significant outcomes.

Learning & Compassion

We aim to provide and participate in an unbiased process with **open hearts and minds**. Our work emphasizes shared learning, open dialogue, and direct communication.

Transparency & Accountability

We remove ambiguity by building **a culture of cooperation, trust, and accountability** so our Partners can succeed, and so everyone can understand how AI systems work.

Diligence & Excellence

Scientific rigor is at the forefront of our **evidence-based approach** with leading AI experts, scientists, and researchers who are engaging to produce impactful knowledge-based outputs.

Investigating Challenges to Diversity in AI

The lack of diversity in the field of AI has been well-documented. As an industry, AI struggles to both recruit and retain team members from diverse backgrounds. But despite widespread awareness of AI's diversity gap, the crisis continues. In July of 2020, PAI formalized its commitment to investigating this persistent problem with **the hire of our first Diversity and Inclusion Fellow**.

Amid growing calls for action, significant investments have been made into Diversity, Equity, and Inclusion (DE&I) efforts, but there remains a lack of clarity about which initiatives work best. While companies have devoted significant time and money to DE&I recruitment activities in recent years, limited research exists on what happens to women and minorities after they enter technical professions.

Working in partnership with DeepMind, PAI researchers launched a study last fall designed specifically to **investigate high attrition rates among women and minoritized individuals in tech**. Through interviews and questionnaires with individuals working on AI teams who are female-identified or who identify with a minoritized identity, as well as DE&I leaders and managers of all backgrounds, this study is gathering desperately needed information about the factors leading to greater inclusion among AI teams.

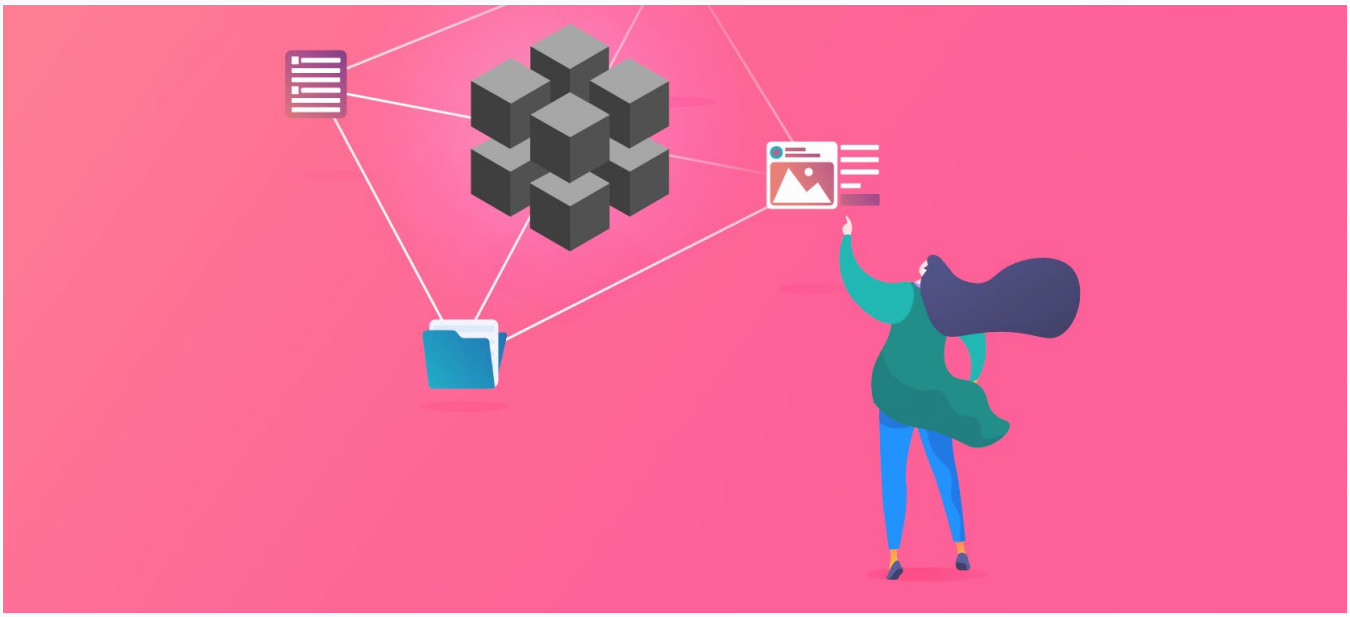
This necessary work has never been more urgent for the AI industry. Low representation of women and Black people in AI may lead to significant racial bias encoded within algorithms. And all organizations that use AI in their work are increasingly aware of the need to actively challenge bias and discrimination in the products they produce.

By gathering actionable insights, PAI plans to release **future products aimed at addressing and improving the issue of attrition of diverse professionals in AI**. Organizations may turn to these products as resources in striving to make a more inclusive environment for people working in AI.

"PAI's Shared Prosperity Initiative Steering Committee has been a valuable forum for debating AI's impact on workers and the economy, and in the face of rapid technology change, to develop collective thinking on the structural changes necessary for a more democratic and just economy."

Andrea Dehlendorf
Co-Executive Director
United for Respect

Convening Across Industries



In service of transparency and accountability goals, PAI hosted **a one-day, in-person workshop around the deployment of "explainable artificial intelligence" (XAI)** in February of 2020. Our paper detailing the takeaways from this workshop was accepted as a poster at the 2020 ICML Workshop on Extending Explainable AI Beyond Deep Models and Classifiers and as a spotlight at the Workshop on Human Interpretability in Machine Learning.

Explainability is often proposed as a way of increasing the public transparency of AI, but current XAI implementations are rarely catered to the public. The XAI workshop stemmed from an interview project PAI conducted in 2019 which found that **the majority of XAI deployments are not for end-users affected by the models, but rather for machine learning (ML) engineers** debugging the models themselves. This workshop aimed to leverage PAI's multi-stakeholder convening capacity to bring together a diverse group of ML developers, researchers, designers, policymakers, and legal experts to explore this gap and outline paths forward for reaching a wider variety of stakeholders with model explanations.

Workshop participants were split into interdisciplinary discussion groups of 5-6 individuals facilitated by a member of PAI staff. Despite there being a multitude of groups bringing their own definitions of "explainability" to the discussions, the focus of these definitions were notably consistant. Most explainability definitions included references to: context, the scenario in which models are deployed; stakeholders, those affected by the models and those with a vested interest in the explanatory nature of the models; interaction, the goal the models and their explanations serve; and summary, the notion that "an explanation should compress the model into digestible chunks."

This exercise **reinforced recent scholarship advocating for situated and contextualized explanations** because abstracted explainability metrics are seen as unlikely to succeed on their own.

"I appreciate that PAI is a convener, able to draw industry, civil society, and academia together for discussions and debates about important issues."

Lartease Tiffith
Senior Manager, Public Policy
Amazon

Taking a Stand Against AI Misuse

Facing an unprecedented global pandemic last spring, public officials were interested in using every tool available, including AI systems, in the fight against COVID-19. Not every instrument, however, was equally appropriate for the job. When former Attorney General William Barr recommended misusing a pre-COVID algorithm to determine potentially life-or-death outcomes for federal prisoners, **PAI released an issue brief explaining the many perils of this path**.

In response to the COVID-19 crisis, Barr issued a memo in March 2020 identifying six factors to consider when deciding which federal prisoners should be prioritized for transfer to home confinement. One of these factors was each inmate's score in PATTERN, an algorithmic tool created to predict federal prisoners' risk of rearrest. Inmates with PATTERN scores above "minimum," Barr wrote, should not be prioritized. Notably, a previous version of PATTERN assigned a minimum score to just 7% of African American males compared to 30% of White males.

In April, PAI published "Algorithmic Risk Assessment and COVID-19: Why PATTERN Should Not Be Used." **This paper detailed why Barr's suggested use of the "PATTERN" risk assessment tool was likely to increase COVID-related racial disparities**.

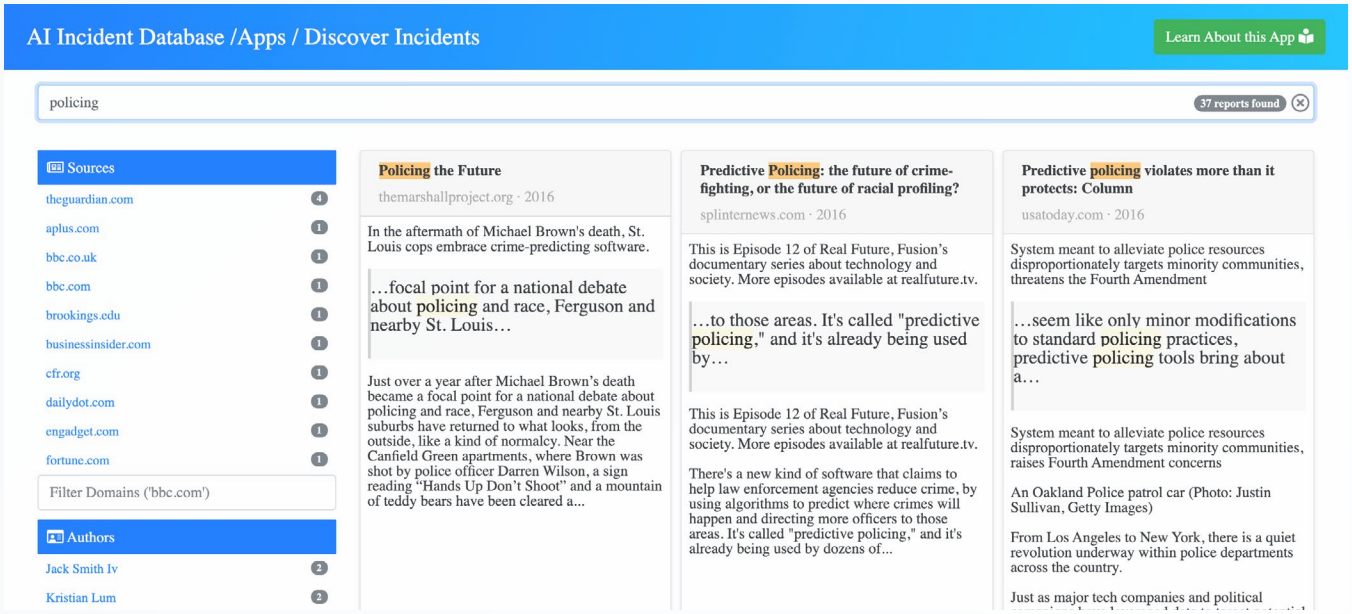
Due to racial bias in the data used to develop, validate, and score PATTERN, PAI explained, using it to guide home confinement decisions would likely to contribute to racial disparities. Furthermore, to the extent that the tool is useful, it is for its designed purpose as a predictor of future arrest – not future criminal activity, much less criminal activity under home confinement during a global pandemic.

While timely, the paper was merely the latest result of **PAI's ongoing efforts to examine questions of racial bias in algorithmic tools** used within the criminal justice system.

"PAI plays a crucial role in facilitating global progress on AI ethics issues. PAI's multi-stakeholder approach – bringing together industry, civil and human rights organizations, academia, and media – is key to driving change in large corporate organizations that ultimately affect billions of lives through AI products."

Michael Spranger
COO
Sony AI

Tracking When AI Systems Fail



In November, PAI publicly launched **the AI Incident Database (AIID), composed of more than 1,000 reports of AI failures** that caused harms or near-harms. Stewarded by a representative of PAI Partner XPRIZE, the AIID serves as a much-needed tool for AI researchers and developers, outlining a wide variety of real-world risks for automated systems.

As a Vice News story covering the database's debut put it, "the platform is being used to document and compile AI failures so they won't happen again."

Authorities like the Federal Aviation Administration have long maintained databases of bad outcomes for the benefit of their respective fields, but no systematic repository of harmful incidents previously existed for members of the AI community. The necessity of such a system has only become more urgent in recent years as AI increasingly enters safety-critical domains like healthcare, criminal justice, and transportation. **To learn from the mistakes of the past, the past must first be chronicled** – a goal the AIID was created to help achieve.

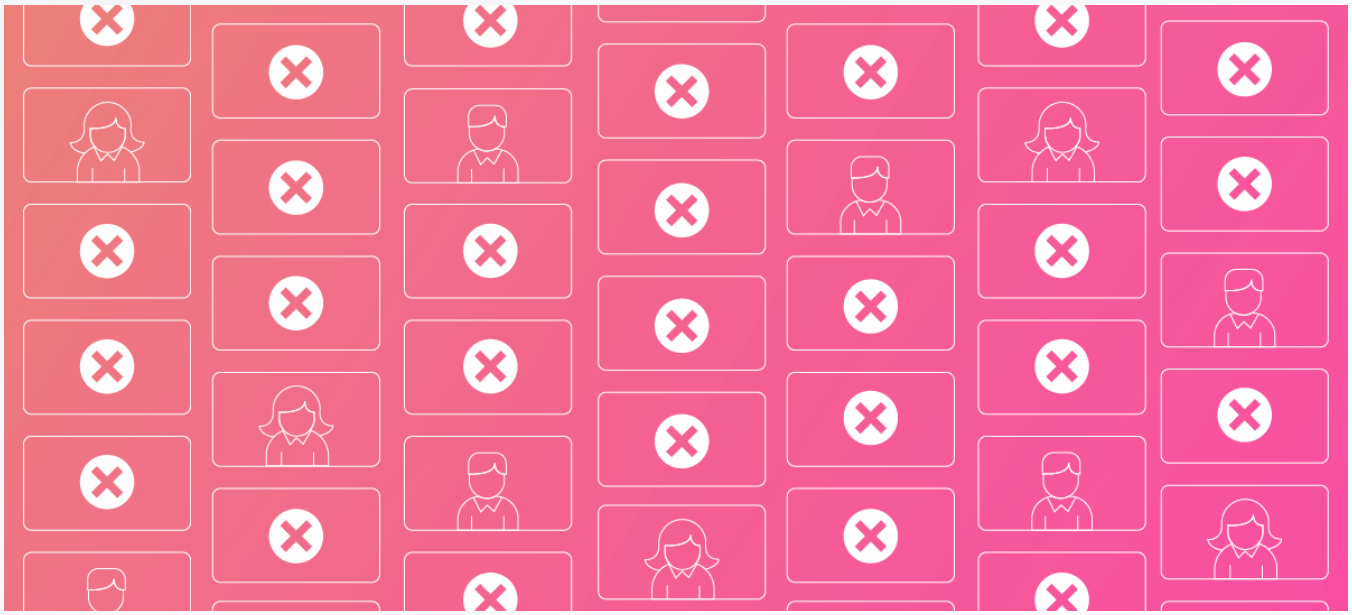
This project sprang directly from the needs of the PAI research community, which found there to be no prior resource offering a comprehensive overview of AI safety and fairness failures. Now, with the crowdsourced AIID, an AI researcher can type "policing" into the database's search field and retrieve dozens of citable incident reports on the topic. Similarly, corporate product managers hoping to anticipate and mitigate risks could search the AIID for "translate" before launching a new translation service. They then might learn about Incident 72, when a social media user was reportedly arrested after his "good morning" message was automatically translated as "attack them."

Created in response to a practical challenge and cultivated through contributions from the greater AI community, the AIID exemplifies **the kind of novel solutions fostered by PAI's collaborative approach to responsible AI**. Solutions that ultimately work to the benefit of not just some, but all.

"It is my privilege to be part of PAI's thoughtful effort in bringing diverse voices and fostering robust debate so that we are deliberate about ensuring that one of the most transformative technologies of our times truly benefits the many, and not just the few."

Rahul Panicker
Chief Research and Innovation Officer
Wadhwani AI

Taking a Methodical Approach to Best Practices



Championing accurate information became critical in 2020, when even minor misapprehensions about COVID-19 could threaten everyone's well-being. Throughout the year, we worked with PAI Partner First Draft to support information integrity, **investigating what works (and what does not) when addressing deceptive content online**.

This collaboration bore its first fruits in June with the publication of "It matters how platforms label manipulated media. Here are 12 principles designers should follow." Drawing from existing academic literature, original research, and interviews with industry experts, **this guide provided a concrete set of principles for decision-makers at social media platforms** seeking to minimize harms.

Amid rising awareness of their role in the spread of misinformation, social media companies have become increasingly proactive in moderating and labeling false content online. At the same time, the real-world impact of these interventions remains insufficiently studied. Platforms are now doing more, but what actions will actually reduce the internet's mis- and disinformation problem?

Through our Media Integrity Issue Area, PAI has begun to answer this fundamental yet often neglected question. The design principles we published in June established a research-based foundation for the responsible labeling of manipulated media. Additional outputs in 2020 provided social media platforms with **specific, immediate recommendations for the automated categorization of manipulated media** and used interviews with end-users to identify the limitations of common intervention strategies.

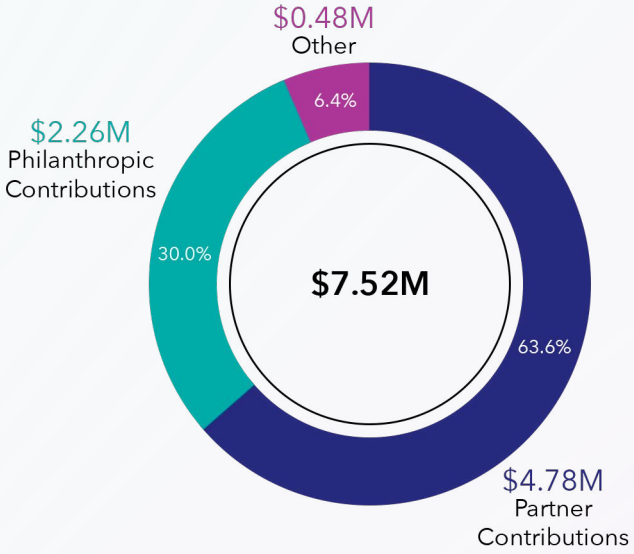
Like so many challenges currently facing the AI community, internet misinformation cannot be solved without inviting all stakeholders to share their varying needs and perspectives. At PAI, these stakeholders come together to work on urgent goals.

"I have been incredibly impressed with PAI's focus on media integrity issues as they relate to platform interventions. The work has been professional, rigorous and much needed."

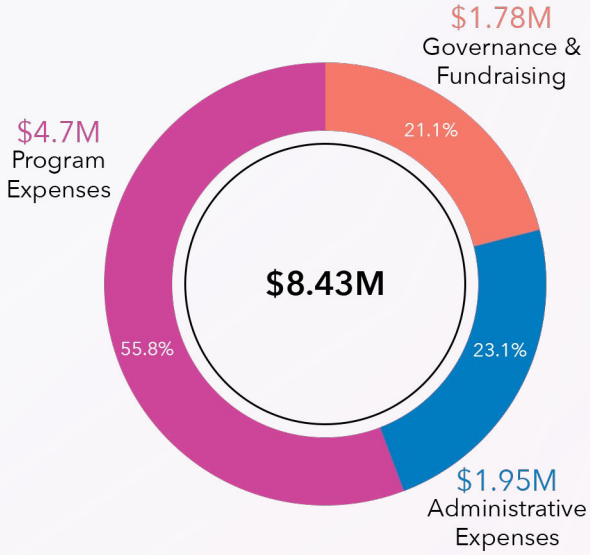
Claire Wardle
Co-founder and U.S. Director
First Draft

FINANCIALS

Revenue



Expenditures



These numbers are projected and have not yet been audited.

Thank you to our philanthropic funders. Your support makes PAI's work possible.

FORD FOUNDATION

KNIGHT FOUNDATION

Luminate
Building stronger societies

MacArthur Foundation

pivotal VENTURES
A MELINDA GATES COMPANY

skoll FOUNDATION

ALFRED P. SLOAN FOUNDATION

THANK YOU

PAI works with numerous organizations and individuals in a variety of sectors. **We thank all of the Partners who form this special community** for their valuable insight and support.

Academic Partners

3A Institute • AAAI • AI4ALL • Allen Institute for AI • Berkeley Center for Law & Technology Policy (BCLT) • The Berkman Klein Center at Harvard University • Center for Human-Compatible AI (CHAI), UC Berkeley • Center for Human Rights Science at Carnegie Mellon University • Center for Information Technology Policy • Center for Internet and Society • Cornell CIS • Duke Reporter's Lab • The Hastings Center • Hong Kong University of Science and Technology, Centre for Artificial Intelligence Research (CAiRE) • Humanity Centered Robotics Initiative, Brown University • Insight Centre for Data Analytics - University College Cork Ireland • Leverhulme Centre for the Future of Intelligence (CFI) • Markkula Center for Applied Ethics • MIT Initiative on the Digital Economy • The MIT Media Lab • Oxford Internet Institute • Tufts HRI Lab • University College London Faculty of Engineering • University of Tokyo - Next Generation Artificial Intelligence Research Center • USC Center for AI in Society • Vision and Image Processing Lab at Waterloo University

Nonprofit Partners

ACLU • AI Forum of New Zealand • AI Now • Alan Turing Institute • American Psychological Association • Article 19 • Association for Computing Machinery • Berggruen Institute • Business for Social Responsibility (BSR) • Carnegie Endowment for International Peace • Carnegie-Tsinghua Center for Global Policy • Center for Data Innovation • Center for Democracy and Technology (CDT) • Chatham House • CIFAR • Code for Africa • Data & Society • DataKind • Digital Asia Hub • Digital Catapult • Electronic Frontier Foundation (EFF) • First Draft • Fraunhofer IAO • Full Fact • Future of Humanity Institute • Future of Life Institute • Future of Privacy Forum • G3ict • GLAAD • Human Rights Data Analysis Group (HRDAG) • Iridescent • The Joint Center • Longpath Labs • Meedan • Mozilla Foundation • New America Foundation • Optic Technology - Human Technology Foundation • Organization United for Respect (OUR) • PolicyLink • Samasource • Shift7 • Software.org - The BSA Foundation • Tech Policy Lab • Underwriters Laboratories, Inc. • UNICEF • United Nations Development Programme (UNDP) • Upturn • Wadhwani Institute for Artificial Intelligence (Wadhwani AI) • Wikimedia Foundation • WITNESS • Women in Machine Learning & Data Science • XPRIZE

Industry Partners

Accenture • Adobe • Affectiva • Amazon • Apple • DeepMind • Essence Global • Facebook • Google • IBM • Intel • McKinsey & Co. • Microsoft • OpenAI • Samsung • Softbank • Sony

Media Partners

BBC • Canadian Broadcasting Corporation (CBC) • The New York Times

PARTNERSHIP ON AI

115 Sansome St. ste. 1200
San Francisco, CA 94104
www.partnershiponai.org