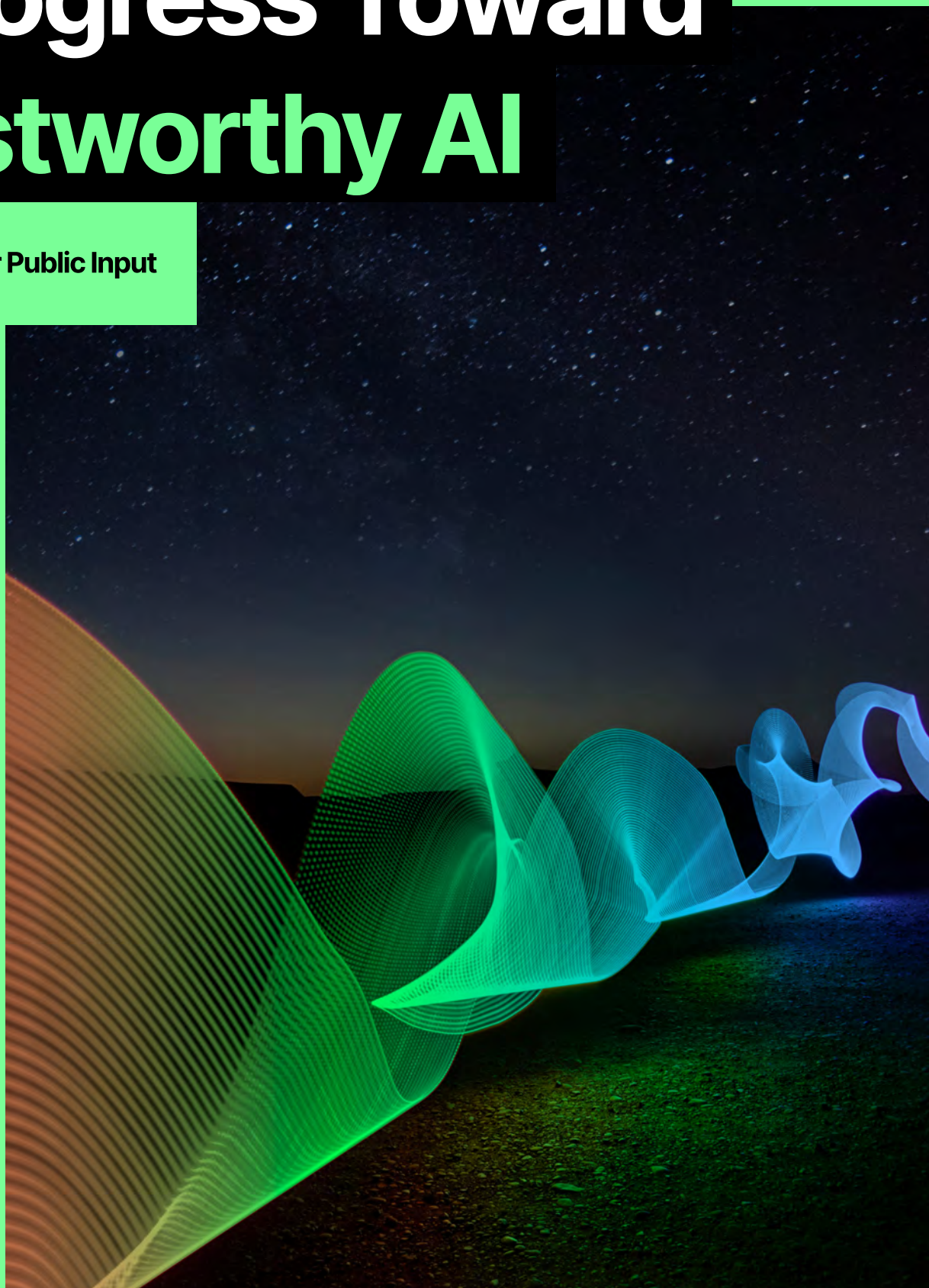
Accelerating Progress Toward Trustworthy AI

Version 0.9 - For Public Input



moz://a

Status update on our 2020 paper "Creating Trustworthy AI" and next steps to promote openness, competition, and accountability in AI.

Authors: Mark Surman, Ayah Bdeir, Lindsey Dodson, Alexis-Brianna Felix, Nik Marda **Contributors:** J. Bob Alotta, Ashley Boyd, Ian Carmichael, Moez Draief, Max Gahntz, Linda Griffin, Stephen Hood, Saoud Khalifah, Santiago Martorana, Mohamed Nanabhay, Alondra Nelson, Kasia Odrozek, Becca Ricks, Victor Storchan, Udbhav Tiwari, Imo Udom, Suba Vasudevan, Claire Woodcock.

moz://a

Contents

Read Me

Mozilla's work on AI is not new – we've been funding, building and advocating for trustworthy AI approaches for years. In 2020, we published a white paper that outlined our vision for trustworthy AI in a nascent moment for the technology. Since then, a lot has changed. Today, AI is increasingly powerful and pervasive in our society, and its promise and perils are becoming even more apparent. While there are a growing number of individuals and organizations working on making AI more trustworthy, the scope of the challenge is also continuing to grow. And while we've made progress on advancing guardrails for AI systems, the AI ecosystem has also become increasingly closed and concentrated.

Given the rapid changes in AI recently, we felt there was a need to take stock of the progress we've made so far and the work that is left to do. This new report provides an update on work to date in the four strategic areas we outlined in our 2020 paper, and it maps key initiatives happening in these areas – both at Mozilla and across the ecosystem. Importantly, the paper also emphasizes our evolving focus on the centrality of open source in the development of more trustworthy AI. We hope this report will be both a guidepost and map — helping readers to articulate their own strong message on trustworthy AI, build and invest in a better future for AI, and find opportunities to collaborate with others in the AI ecosystem. In turn, we believe our collective work can set us on a path to truly advance openness, competition, and accountability in AI.

We invite your input on the report and your feedback on the state of the AI ecosystem more broadly. Through your comments and a series of public events, we will take feedback from the AI community and use it to strengthen our understanding and vision for the future of trustworthy AI. Please email us at AIPaper@mozillafoundation.org to provide any input on the report and/or to highlight your favorite examples of AI being used in ways that build trust and improve people's lives.

01 Executive Summary

“

In a world where AI is touching every sector and facet of society, we need structural changes that tackle the root causes of today’s AI harms and unlock the positive benefits of AI.

”

AI in our world today →

We're at an inflection point for AI, and for society at large. The technology is unlocking huge societal benefits, ranging from AI-powered drug discovery and climate solutions to productivity gains for individuals and small businesses. While these benefits are profound, the harms from AI have also never been more pressing. We're seeing AI being used in ways that make it easier to deceive and harass people on social media, perpetuate bias in the criminal justice system, and extract sensitive information from people's online activity.

Today's AI ecosystem is structurally flawed in ways that prevent us from realizing the full potential of AI, while also allowing AI harms to go unchecked. We know that many of AI's innovations and applications have been fueled by open source and open science; for example, Google's influential transformer paper and TensorFlow framework were made widely available, which supported many AI innovations across sectors. But now, many big tech companies are vilifying open and competitive approaches to AI in favor of their own proprietary AI models and lucrative cloud computing businesses. This is making it harder to compete in the AI ecosystem.

We know from other industries that competition is vital for spurring research and development, creating cheaper and safer products, and invigorating investment and job creation. A lack of openness and competition also makes it harder to promote accountability in AI, as it reduces independent research and collaboration, inhibits scrutiny from the public and regulators, and increases market barriers for new players focused on creating responsible AI. In a world where AI is touching every sector and facet of society, we need structural changes that tackle the root causes of today's AI harms and unlock the positive benefits of AI. **That's how we get to trustworthy AI: tech that is built and deployed in ways that support accountability and agency, and advance individual and collective well-being.**

A familiar story for Mozilla→

At Mozilla, we're deeply familiar with this situation. At the dawn of the commercial internet in the late 1990s, Microsoft was on the brink of monopolizing the market for web browsers, threatening to lock in users, stamp out competitors, and stifle innovation online. The internet was poised to transform society, but access to the internet was increasingly being controlled by one entity. In response, Mozilla created the open source Firefox browser that added much-needed competition in the marketplace, raising the standard for privacy, security, and functionality across the industry. In the 25 years since, we have continued fighting the power of big tech on the internet through our products, investments, and advocacy.

That's why we were concerned when we saw a similar pattern emerging in the AI ecosystem over the last decade. In 2020, we articulated a vision for trustworthy AI that highlighted many of the issues we saw with the AI ecosystem. We outlined four levers that we could pull to achieve trustworthy AI at scale, and mobilized the Mozilla community toward pulling these levers.

Since then, Mozilla has been actively building, investing, and advocating to push even further down this path. We're investing in new technology and new products to demonstrate trustworthy AI principles in action, offering consumers more choice and control. We're educating policymakers around the world on the existing risks and benefits of AI to shape smarter regulations and fairer market dynamics. We're continuing to rally like-minded builders, researchers and activists to drive consensus around responsible AI development and fund initiatives to make AI more trustworthy. We're helping consumers be more critical in choosing AI products, and encouraging lawmakers to prioritize openness and accountability in AI policy.

Progress and next steps towards trustworthy AI→

In our original paper, we proposed four key levers for advancing the development of more trustworthy AI: **(1) changing AI development norms**, **(2) building new tech and products**, **(3) raising consumer awareness**, and **(4) strengthening AI regulations and incentives.** This report outlines where we've made the most positive progress within each lever, and where there is still more work to be done.

Key Takeaways →

01

Norms

Changing AI development norms

The people that broke the internet are the ones building AI.

Big tech companies and the AI startups they back currently dominate AI. They have created opaque, centralized AI models using our harvested data. Luckily, there is a growing wave of startups, builders, educators, scholars, researchers, and civil society leaders focused on shifting these norms, with a focus on building open, transparent, and trustworthy AI.

02

Products

Building new tech and products

More trustworthy AI products need to be mainstream.

Over the last 18 months, black box generative AI tools have entered the mainstream of business and public consciousness. At the same time, dozens of startups and research projects have sprung up to build open source models, auditing tools, and data platforms that offer a different path. While not yet mainstream, these are the seeds of a better AI ecosystem.

03

Consumers

Raising consumer awareness

A more engaged public still needs better choices on AI.

Consumers are starting to pay attention to AI's impact on their lives. Workers — from delivery drivers to Hollywood writers — are pushing back on how AI affects their livelihoods. However, we have not yet seen a wave of mainstream consumer products that give people real choice over how they interact with AI. This is a key gap in the market.

04

Policy

Strengthening AI regulations and incentives

Governments are making progress while grappling with conflicting influences.

As policymakers move to regulate AI, they are confronted by conflicting messages, especially from industry. Don't regulate, says one camp. Limit control over cutting-edge AI to a few companies, says another. Some regulators are taking a third way, listening to less-prominent voices in industry, academia, and civil society arguing for a balanced approach.

“That’s how we get to trustworthy AI: tech that is built and deployed in ways that support accountability and agency, and advance individual and collective well-being.”

This report shows that we’ve made meaningful progress since 2020, and that there is still much more work to be done. It’s time to redouble our efforts and recommit to our core principles, and this report describes how. It outlines Mozilla’s ongoing work to shift the narrative on AI, make open source generative AI more trustworthy and mainstream, and empower consumers with real choices on AI. It highlights how we’ll continue investing in the trustworthy and open source AI ecosystem, and help lawmakers develop and roll out pragmatic AI regulation. And, it calls on builders, consumers, policymakers, advocates, and investors alike to leverage their respective positions to push the AI ecosystem in a better direction.

It will take all of us, working together, to turn this vision into reality. There’s no time to waste — let’s get to work.

02 Introduction: AI Challenges, Risks, and Opportunities

Heightened public attention on AI →

Investments and advancements in AI have been ongoing for decades, but the technology's transformational benefits and potential for long-lasting harm have exploded in the last three years. The release of generative AI systems based on large language models (LLMs) like ChatGPT, Bard, and Stable Diffusion have captured the public's imagination, allowing everyday people to interact with AI systems using natural language for the first time. Widespread consumer awareness has sparked an AI gold rush for big tech, entrepreneurs, and investors, with new players and big tech incumbents battling to dominate the fast-growing market. At the same time, more people are calling out how AI models can be abused, biased, opaque, and harmful at scale. The media has struggled to discern AI hype from reality. The public has mixed feelings about AI, with some fearing for their jobs. Artists have sued over alleged copyright infringement in AI training data. And activists, academics, and technologists are having intense debates over which AI dangers need to be addressed first, and which AI development approaches are the safest.

"At Mozilla, we fall into the camp of AI realists, who are cautiously optimistic about AI's positive potential and dedicated to solving present day harms."

Heightened awareness of AI risks is a good thing. Without a well-informed public pushing for more transparency and accountability, large tech companies will use AI to pursue profit over all else, which harms people by further consolidating their market power. If they do, we could miss our chance to shape an AI landscape that benefits people everywhere, and not just increases profit margins for Silicon Valley players.

Much of the debate about the future of AI has focused on two positions: unbridled optimism and existential fear. One side argues that virtually all AI, and all technologies, are universally good, and should not be "restricted" by regulation or other risk mitigation approaches. The other side argues that AI poses an existential threat to humanity, and must be constrained in ways that can both limit current AI benefits and exacerbate existing AI risks.

However, there is a third school of thought that offers a more nuanced, practical perspective on the risks and benefits of AI. At Mozilla, we fall into the camp of AI realists, who are cautiously optimistic about AI's positive potential and dedicated to solving present day harms like AI bias, discrimination, and job loss. This camp is not new. Many individuals and civil society organizations have been promoting this perspective for years. Alongside the broader open source community, and other foundations, think tanks, researchers, and activists, we believe that addressing AI's problems today with openness and transparency will serve the greater good for society and the economy while mitigating both everyday and catastrophic risks.

Introduction: AI Challenges, Risks, and Opportunities

To inform our collective next steps, we lay out the five core challenges that we see in today's AI landscape:

Many people got the AI story all wrong.

Over the past year, the story of AI pitching a battle between the optimists and the doomers sucked attention away from more pragmatic and thoughtful approaches to AI. Out of the media limelight, AI realists were rolling up their sleeves to unlock the huge benefits of AI while also tackling tough questions about social ills and closed markets. This can — and should — be the story we're all focused on.

Big tech and closed models are dominating the field.

The increasing capabilities and adoption of AI have made it even more important to enable competition and market access, independent research and public scrutiny of AI systems, and more room for new players to build trustworthy AI products. Over the past few years, the AI ecosystem took a radical swing in the direction of closed technology — leading AI companies stopped publishing papers and started selling access to APIs. At the same time, big tech invested heavily in players like OpenAI and Anthropic as a way to control the field and bolster their own cloud computing businesses. This is making it harder to advance the open approaches that are vital to creating a better AI ecosystem for everyone.

Open source generative AI hasn't hit the mainstream.

Over the last year, a huge wave of open source generative AI models was released — from Llama to Falcon to Mistral. While these new models are gaining steam and offer huge promise, there is still a long way to go before they are easy to use and easy to trust. Open source generative AI won't hit the mainstream until we tackle these issues.

There are still foundational issues in AI development.

The root cause of many AI harms can be traced to foundational aspects of the AI ecosystem, including the population that's building AI and the ways they're collecting and labeling data. For example, because most LLM training datasets are built using widely-available data from across the web, the systems reproduce biases and stereotypes that cause real-world harm. The tech industry also still lacks diversity, which means valuable perspectives are left out of AI development, preventing these models from reflecting the breadth and depth of the human experience.

Policymakers are moving, but the tech and harms are moving faster.

In 2023, progress on the European Union AI Act and the release of the U.S.' Executive Order on Safe, Secure, and Trustworthy AI have shown that policymakers understand the urgency of firmer political action on AI. However, the rapid growth of AI and its associated harms are moving much faster than the development and rollout of new regulations. In the meantime, companies like OpenAI, Meta and Google are having meaningful impacts on the economy just one year after ChatGPT's release.

> We have come to believe that open source approaches and broad market access are key ingredients for promoting agency and accountability in the AI era.

“Building trustworthy AI is both urgent and complicated, and no one person or organization can tackle all of these risks alone.”

This watershed moment in technology history is the best opportunity for the AI realists to influence the direction of the industry, driving us toward a better technological future for all. Building trustworthy AI is both urgent and complicated, and no one person or organization can tackle all of these risks alone. That’s why we emphasized the need for collaboration across the entire ecosystem in our original 2020 Creating Trustworthy AI paper, and why we’re working alongside others on the products, research, and policy needed to advance trustworthy AI.

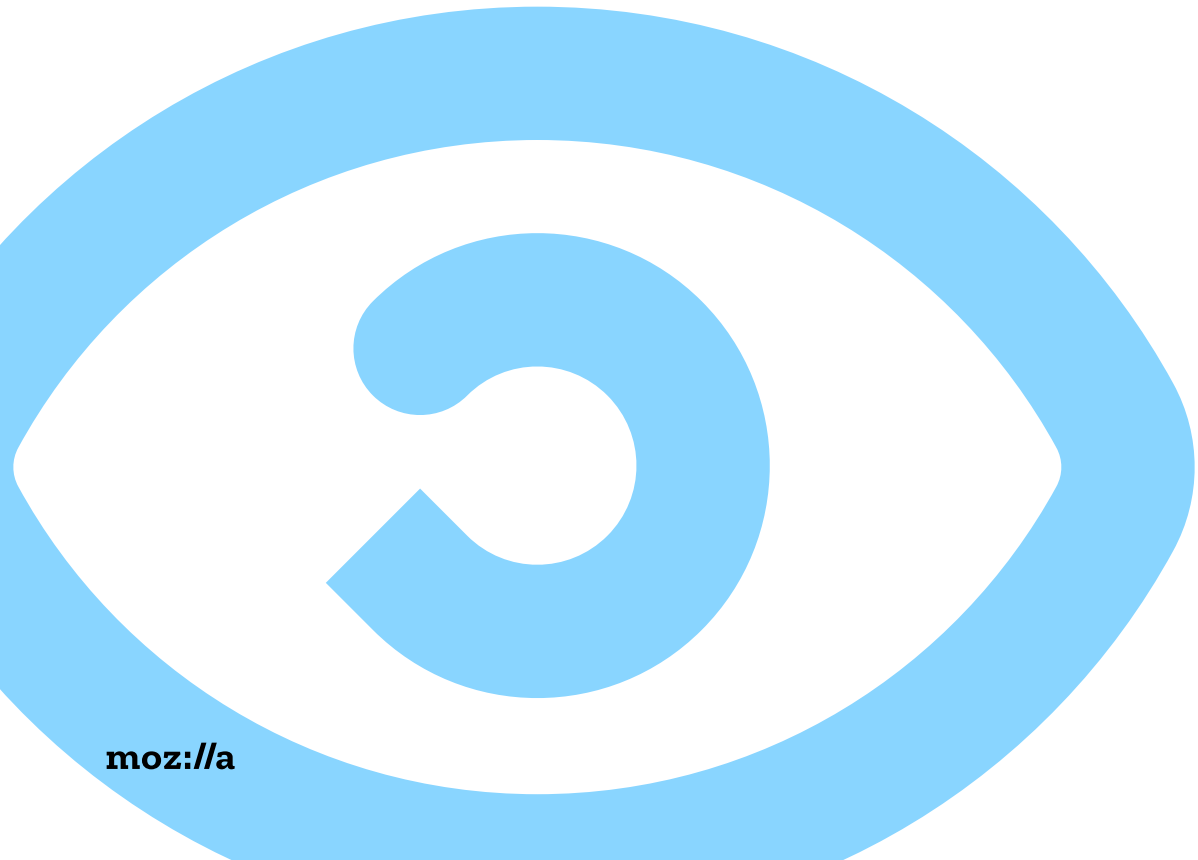
Advancing openness and competition →

In our 2020 paper, published well before the conversation about AI reached its current fever pitch, our team at Mozilla outlined the importance of creating AI that people can trust, with agency and accountability at the center. Agency allows users to regain control over their internet experiences, with a focus on privacy, transparency, and well-being. Accountability means companies are held to account when their AI systems cause harm through discriminatory outcomes, abuse of data, or unsafe practices.

Since then, we have come to believe that open source approaches and broad market access are key ingredients for promoting agency and accountability in the AI era. We’ve noticed that AI is increasingly being built in a closed ecosystem by the same companies that have already damaged both the internet and society over the last 20 years. Many of AI’s most exciting innovations flow from open source and open science, but researchers are increasingly moving away from open publishing as their corporate funders prioritize selling access to cloud services. These same companies are now vilifying open source approaches to AI in favor of their own proprietary models. Newer players have entered the market with financial backing from the big tech incumbents, and are advocating for limitations on who can access or build the most powerful AI systems, citing potential security risks as a part of their critique of open source AI. However, closed models can also be abused by bad actors and deployed by ill-equipped developers, and openness is actually a key ingredient towards safety, security, and accountability. We cannot accept that a black box approach, kept in the hands of just a few companies, is the only safe and sensible path forward.

While open source alone will not cure all of AI's problems, when done responsibly, it can foster an environment of innovation, transparency, and community building. Coupled with competitive markets, it can help ensure advancements in AI are accessible to all, contributing to the collective knowledge base and allowing diverse perspectives to shape and govern the technology. That's why we're funding, building, and collaborating with partners who are working to make open source AI trustworthy, commercially successful, and useful for humans everywhere. We have a deep history of leveraging open source for societal benefit, and we're working to do it again for AI.

As this report will show, both Mozilla and the broader ecosystem are making meaningful progress toward trustworthy AI, but more work remains. We know we can't turn the tide alone, which is why this report — and all our work — is about creating a global community committed to trustworthy AI. That's how we build a better future.



03

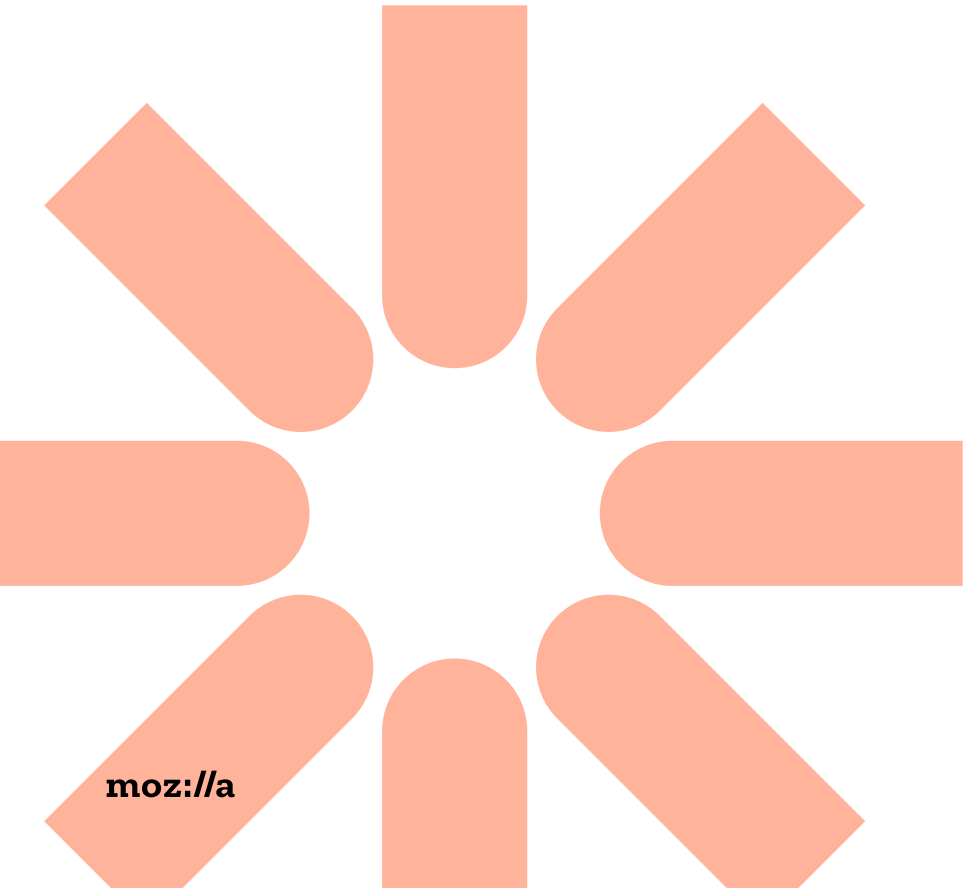
Changing AI Development Norms

Changing AI Development Norms

In our 2020 paper, we called for a shift in the tech industry's norms and guidelines around how AI is built. We highlighted the need for more diversity among the stakeholders involved in designing AI as a key condition of trustworthiness. Today, we're seeing the impact of that lack of diversity come through both in the datasets used to train LLMs, and in who gets to set AI development norms.

AI models are trained on reams of data from across the internet, but because the internet is not universally accessible, the data we use is inherently incomplete. As a result, LLM outputs are limited by the languages and content that are most prevalent online, including hate speech and other harmful posts from the dark corners of the web.

Additionally, if only a small group of people in Silicon Valley are building AI and developing the trust and safety practices that will shape the industry's future, AI products will lack the nuance of other cultural perspectives and lived experiences. That has real consequences for people and communities historically shut out of tech, who will feel the impact as AI use spreads.



Positive Progress→

Though the tech industry is still overwhelmingly white and male, women and people from underrepresented communities have spent the last three years pushing to influence the trajectory of AI development.

One example is the work of Dr. Timnit Gebru, who was fired from her role as a co-lead of Google's Ethical AI team in 2020 for raising concerns about bias in an early LLM version. She has since become an advocate for diversity in tech, speaking out about the centralized market power of large corporations building AI systems that impact the entire world. A year after her firing, she founded the Distributed Artificial Intelligence Research Institute (DAIR), "a space for independent, community-rooted AI research, free from big tech's pervasive influence." Since its inception, DAIR has focused on elevating the voice of marginalized people in research on the harms associated with AI technology. The organization also focuses on building community to accelerate the creation of technologies for a better future.

We also acknowledge the evolution of the ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT), which has made a deliberate effort to increase conference participation from underrepresented groups, expanding the range of voices included in discussions that will shape the future of the industry.

At Mozilla, we've provided financial backing for our Trustworthy AI Fellows, who are addressing a wide range of issues, from racialized algorithmic systems on dating apps to the impacts of AI technologies on rural communities. Several of our Fellows have since been recognized for their work on a global stage, including Inioluwa Deborah Raji and Abeba Birhane, who were named to the 2023 TIME100 AI list for their work on open source algorithmic auditing tools. Birhane's work has challenged the idea that ever-larger AI models will solve the problem of toxic or biased outputs, finding that "as datasets scale, hateful content also scales." Mozilla Ventures also invested in Lelapa AI, which aims to unlock the immense potential benefits of AI technology that has historically excluded African languages.

There is also incredible demand for talent with trustworthy AI expertise, and a very limited number of people who currently have that expertise. Both the tech industry and academia will need to make significant investments in AI education to meet the need. To cultivate a generation of builders with trustworthy AI training, we've partnered with universities on the Responsible Computing Challenge (RCC), which focuses on curricula that empower students to think about the social and political context of computing. RCC has awarded $2.7M in funding to 33 institutions across three continents.

“

AI will impact life around the world, so it’s crucial to have a wider set of voices involved in its design and deployment.

”

Work to be Done→

Widely-accepted industry guidelines on how to build AI responsibly are still in flux. For example, Stanford's Center for Research on Foundation Models released a Foundation Model Transparency Index (FMTI), which aimed to assess the level of transparency for 10 of the top AI foundation model developers. Indexes like this have great promise — and also come with their own limitations. The Stanford analysis quickly came under scrutiny as critics called out the FMTI's shortcomings and bias toward closed models.

We're also seeing a concerning backslide on efforts to diversify major tech companies building AI. Following the murder of George Floyd and subsequent racial justice protests in 2020, leaders across corporate America made commitments to hire, promote, and retain more people of color. That effort started to bear fruit, but the trend could be under threat now that special interest groups are piling in and leaders like Elon Musk are railing against DEI. In 2023, Google and Meta were among the companies who cut back on DEI spending.

We must also consider the lack of regional diversity in AI development. The focus on western-centric efforts obscures the work of builders in areas with less-developed tech sectors, such as South America and the African continent, as well as the exploitation of workers who label data in countries like India and the Philippines. AI will impact life around the world, so it's crucial to have a wider set of voices involved in its design and deployment. The U.N. estimates that nearly 2.6 billion people have no access to the internet, so data used to train AI models is not representative of one-third of the global population's experiences. Without more input from a diverse set of people, emerging guidelines and best practices may not align with the cultural and economic needs of other regions.

At Mozilla, we're continuing our commitment to building and supporting diverse, equitable, and inclusive internal teams. But we know that a broader transformation is needed: engaging a diverse set of stakeholders in shaping AI's future is a core part of Mozilla's programmatic strategy, informing what we fund and where we work. Most notably, the Africa Innovation Mradi has convened and funded AI builders on the continent since 2020, promoting models of innovation grounded in the unique needs of users in Africa.

04 Building New Tech And Products

In 2020, we emphasized the need for more foundational technologies to emerge as trustworthy AI building blocks for developers. As the AI landscape is becoming more closed, it's increasingly critical that these building blocks align with the principles of openness, competition, and accountability.

Open source approaches do not inherently result in trustworthy AI and can be captured and appropriated by corporate interests. However, opening up AI tools to public inspection, modification and remixing is a fundamental first step toward more accountability and agency. Openness can also play a significant role in promoting a fairer AI market, as a healthy open source ecosystem makes it easier for smaller players to compete with incumbents.

We've already seen big tech companies make billions of dollars by stockpiling as much user data as possible, and they are keen to apply that strategy to their AI offerings. That's why the emergence of alternative business models for consumer technologies is a key condition for more trustworthy AI. To promote competition, investors must also support the entrepreneurs pioneering new AI tools and business models that center human agency and protect people's privacy.

Finally, we need to increase development and adoption of tools that can help make AI systems more accountable. For example, privacy-enhancing technologies like federated learning have the potential to reduce risks throughout the AI lifecycle, and help make AI more responsible. Other performance and risk assessment tools can also help developers make AI systems more trustworthy.

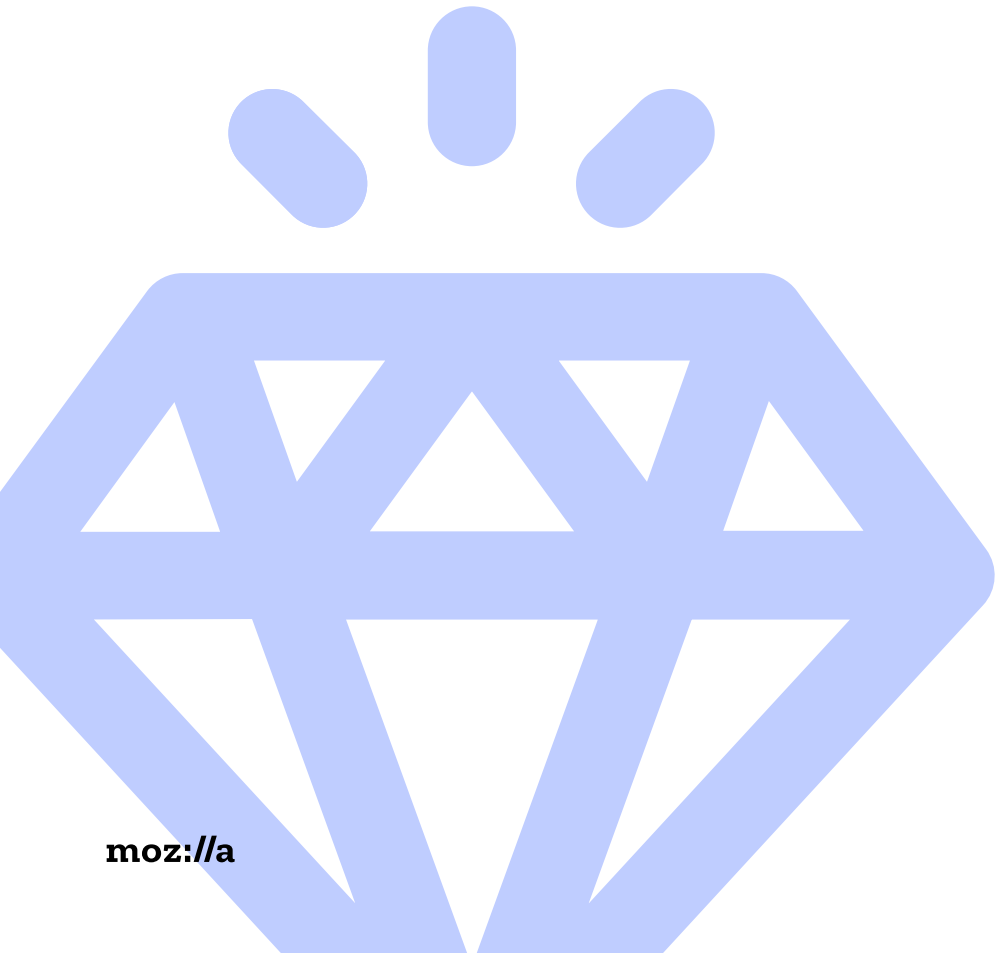
Positive Progress →

Proprietary models got a head start in the latest phase of the AI race, but a range of open source LLMs and resources are emerging to counter them.

Across the AI ecosystem, Meta's family of LLMs has gotten most of the attention in discussions of open source AI, but there are hundreds of other models — like EleutherAI's GPT-NeoX-20B, Hugging Face's BLOOM, the Technology Innovation Institute's Falcon-180B, and Mistral's Mixtral 8×7B — that better reflect open source values. Hugging Face, a platform and community aiming to "democratize good machine learning," is amplifying these models through its Open LLM Leaderboard. By tracking, evaluating, and ranking open AI models submitted by the community, the company is giving small teams and individual developers a vetted foundational resource for building more transparent products.

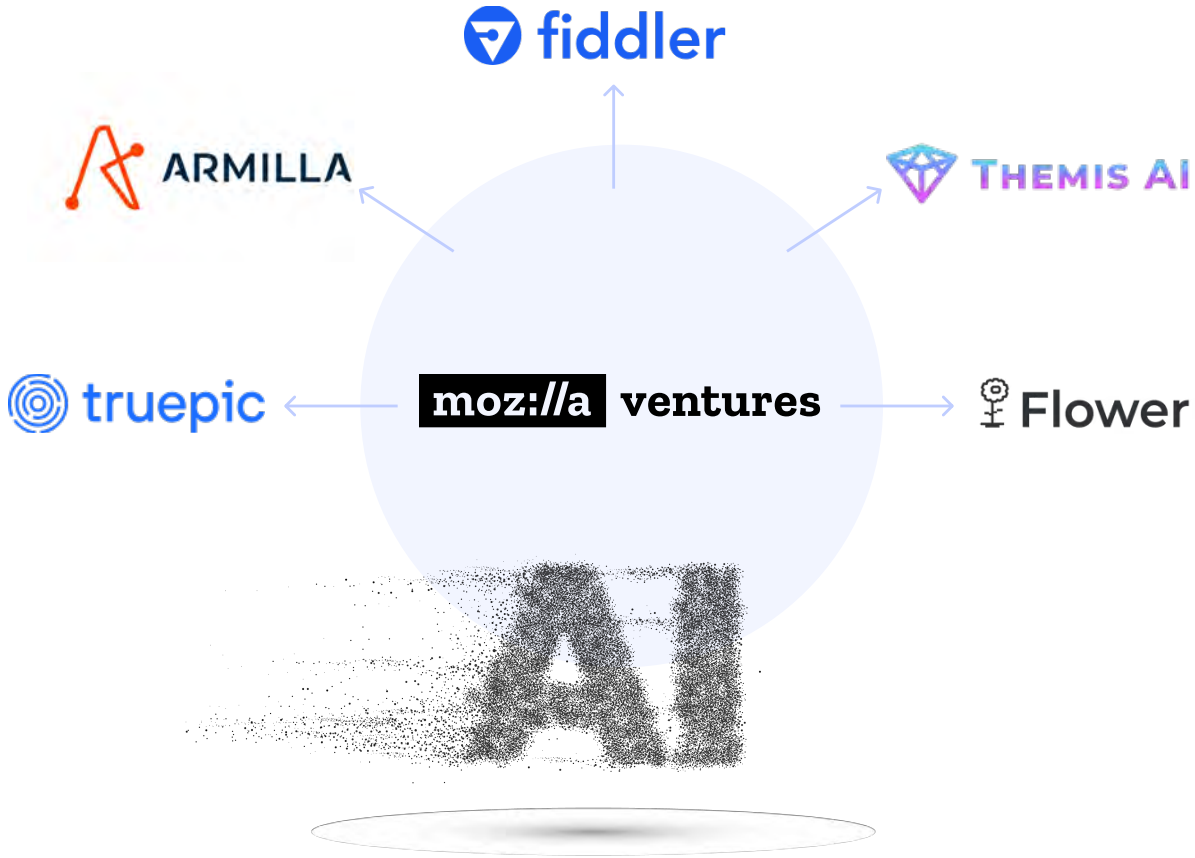
We've also been contributing to this ecosystem at Mozilla. In March 2023, we invested $30 million to create Mozilla.ai, a startup and community dedicated to making this fast-growing field of open source AI models more trustworthy and useful. The company is in the early stages of developing a more communal vision for human-AI collaboration. This includes building small, specialized language models (SSLMs) that can be used to fine-tune models according to knowledge from subject matter experts, while ensuring those experts can tailor the models according to their specific needs. We believe tools like these will help make AI systems more accessible, trustworthy, and useful.

We're also actively developing and investing in open source building blocks that offer more agency and make AI more "local." As part of a broader goal to make AI accessible to everyone through open source, Mozilla released llamafile, an open source, versatile, single file LLM that makes it dramatically easier to run and distribute LLMs on a local machine like a laptop. We've expanded our work on projects like Common Voice — the world's largest multilingual, open-source voice data corpus – which now contains over 100 language data sets, including many local languages not supported by big tech players. Common Voice is the Mozilla Foundation's flagship initiative to mitigate bias in AI by democratizing voice tech for all, and was recognized as a digital public good by the UN-backed Digital Public Goods Alliance initiative.



Mozilla is also working on a number of documentation resources designed to help developers build AI responsibly. This includes our AI Guide, a community-driven collection of open source resources on topics like AI basics and how to choose machine learning models. It also features a set of notable projects from the AI community for insight and inspiration. For example, it includes our Open Source Audit Tooling (OAT) project, which aims to help developers, researchers, and policymakers understand the AI auditing landscape.

We've also invested heavily in the development of the trustworthy AI ecosystem. Mozilla Ventures has already invested $4M in early-stage startups with a focus on trustworthy AI, including Themis AI, Fiddler AI, Armilla AI, Truepic, and Flower. The Themis team has built a software framework to help machine learning models recognize when they are delivering unreliable outputs, while Fiddler AI is building trust into AI by offering observability and explainability tools. Truepic is building content authenticity technologies that can help stop the spread of misinformation via AI-altered images. And, in 2023, Mozilla Foundation committed $1.3 million in technical funding to support trustworthy AI projects through the Mozilla Technology Fund, the Data Futures Lab, and related grant initiatives.



“

Open source AI models are gaining momentum, but they won't go mainstream until they are easier to use, more effective, and more trustworthy.

”

Work to be Done

Open source AI models are gaining momentum, but they won't go mainstream until they are easier to use, more effective, and more trustworthy. To get there, the open source community must focus on making these critical AI building blocks as helpful and successful as possible so they become more relevant in the market. When the barriers to building better AI tools come down, the open source approaches will improve and the trustworthy AI ecosystem will grow.

One of the biggest barriers to open source AI development is the tremendous amount of computing power needed to build and train LLMs. Chip company NVIDIA is fielding record demand for its graphics processing units (GPUs), which can cost as much as $30,000 each. Training a model like GPT-4 requires thousands of those chips, making it prohibitively expensive for small teams and individual developers to build out their own AI infrastructure. If deep-pocketed entities like Microsoft, Google, and the companies they back are the only ones that can afford enough GPUs to train their models, more transparent and trustworthy AI systems will never get off the ground. Governments can help by funding AI compute capacity for public research projects and local startups, as they have begun to do in the U.S., U.K., and the EU. Developers are also working on making it easier to build AI systems using AMD chips, as with our recent update to llamafile.

Another challenge to building new tech and products is a lack of clarity around what "openness" means in the context of AI. The open-source community has yet to reach consensus on a concrete definition of open source AI, or on the right guardrails for releasing AI models to the public. This is critical, as openness alone will not lead to the creation of trustworthy models or mitigate their risks. In September 2023, French startup Mistral AI released its own open source LLM called Mistral 7B, which it claimed was more powerful than Meta's LLaMA-2. However, researchers quickly raised alarms about the system's lack of content moderation filters, which allowed users to prompt the system for bomb-making and self-harm instructions. Other models have built-in security measures to prevent chatbots from answering similar questions, but Mistral's founder stated that safety is the responsibility of developers of AI applications, not the companies building the LLMs.

To tackle these challenges, Columbia University and Mozilla are collaborating on a series of workshops in 2024 to map the different dimensions of openness in AI. The project will engage individuals and organizations with longstanding involvement in open source to build a broad coalition that can stand up to big tech and encourage builders to responsibly open more of their AI development. With this and other efforts, we'll continue leading the way on defining and developing open source tools and systems that are safe, accessible, and transparent.

05 Raising Consumer Awareness

> A well-informed public is a crucial piece of the AI accountability puzzle.

Raising Consumer Awareness

A key lever in our 2020 paper was generating public demand for more trustworthy AI products. This includes both everyday consumers and the civil society organizations that educate and advocate for their best interests. A well-informed public is a crucial piece of the AI accountability puzzle. When a critical mass of users pushes back on questionable practices, large tech companies have no choice but to make changes to address their concerns. Their bottom lines depend on it.

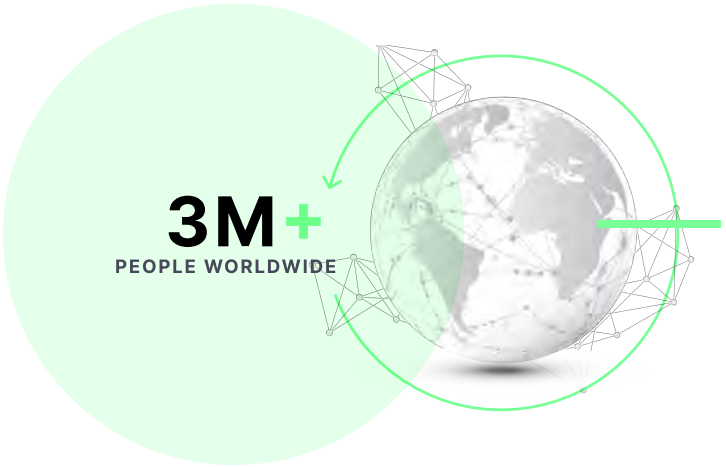
Positive Progress→

ChatGPT's fast rise and extensive media coverage have made AI a mainstream topic. Companies across industries are experimenting with it, which means millions of people are using it in their workplace. As a result, risks that researchers have been warning about for years are in the spotlight. Recent surveys have found that more than three-quarters of consumers are concerned about misinformation resulting from AI, and less than 40% said they believe it's safe and secure. Years of advocacy and policy developments related to privacy, misinformation, and tech platform accountability have created a more informed, skeptical and opinionated public, which is critical for the AI era.

Public opinion around AI is taking shape within familiar contexts of labor, human and consumer rights. Movement leaders have accelerated their understanding of how AI will impact their constituencies, and are shaping attitudes and expectations. For example, AI concerns were a central part of 2023's Hollywood labor strikes. After several months of work stoppages that brought the film and TV industries to a halt, screenwriters and actors secured AI-related concessions in union contracts covering hundreds of thousands of employees. The agreements don't completely prohibit generative AI, but they do place guardrails around how studios can use it, allaying fears about writers' room cuts and AI-generated likenesses. This early victory bodes well for future labor organizing efforts in other industries.

Other work from advocacy organizations has focused on educating consumers about the dangers of AI and encouraging them to choose more trustworthy technologies when available. Consumer Reports released three short films exploring algorithmic bias in medical devices, mortgage lending, and facial recognition as part of its Bad Input campaign. Documentary films like Coded Bias (2021) were important conversation starters for a mainstream audience on topics like misinformation and racial bias in facial recognition algorithms. Dr. Joy Buolamwini's 2023 book Unmasking AI expands on her experiences featured in Coded Bias, and on the founding of the Algorithmic Justice League. Even those with more extreme perspectives on AI have played a key role in raising consumer awareness; for example, in 2023, the Center for Humane Technology gave a widely-watched talk on the existing risks of AI technologies, providing more context on how the race to capitalize on AI can lead to safety failures.

Raising Consumer Awareness



Mozilla has a supporter community of over **3M** people who participate in its trustworthy AI campaigns and education initiatives.

Mozilla's public campaigns and wider advocacy on issues around AI and consumer tech have mobilized over 500,000 people worldwide since 2021, driving meaningful changes to products and industry standards. These campaigns have raised consumer awareness of issues around AI and the tech they use in their everyday lives, from search engines and social media platforms to video doorbells and cars. In July 2023, Slack implemented a blocking feature following a civil society campaign we spearheaded. In September 2023, YouTube announced it would give civil society researchers access to crucial data, following a multi-year campaign by Mozilla. In the same month, the Alliance for Automotive Innovation called for a federal privacy law in the U.S. following public pressure generated by Mozilla's ongoing *Privacy Not Included report series, which most recently focused on privacy issues with connected cars. We also launched a new season of our IRL Podcast in 2023, focused on AI developers bringing responsible products to market.

Beyond advocacy, we're also seeing new consumer products coming to market built on trustworthy AI principles. Established companies like Intel and Adobe, startups like Reality Defender, and research organizations like the MIT Media Lab are working on ways to identify deepfakes, certify image authenticity and fight dis- and mis-information. Twilio has introduced an AI Nutrition Facts initiative, offering a consumer-friendly way to understand how an AI product uses their data. Google's DeepMind group also beta launched SynthID, a tool for watermarking and identifying AI-generated content, in August 2023.

In May 2023, Mozilla acquired Fakespot, a company that protects consumers by using AI to detect fraudulent product reviews and third-party sellers in real-time. Our technology analyzes billions of consumer reviews to quickly identify suspicious activity and then recommend better alternatives to consumers. In late 2023, we launched Fakespot Chat, which uses the power of generative AI to quickly answer shoppers' product questions, saving consumers time and money.

Work to be Done →

Though we've seen positive momentum to give workers more of a say in the introduction of AI tools and systems in the workplace, most casual users of generative AI tools are not thinking critically about whether these systems are trustworthy or not. Since there are few well-known alternatives to popular tools based on closed models, many consumers may feel forced to choose the tools from the companies whose technology is already embedded in their daily lives.

We can't blame consumers for choosing the most convenient tools that appear to be trustworthy, especially when there are limited alternatives. The open source community must continue building AI technology to give people better options. Civil society organizations must keep sounding the alarm on the potential for unchecked AI to cause real-world harm, and funding better alternatives. And regulators must preserve a competitive marketplace with strong consumer protections, giving the broader AI ecosystem the necessary guardrails to thrive.

At Mozilla, we're working to build more trustworthy AI technologies into our own consumer products in the coming years, and are expanding our crowdsourced investigative research into how TikTok's recommendation algorithm works. We'll also need to continue raising consumer awareness of the AI privacy risks and encourage demand for more privacy-preserving approaches used in AI products.

06

Strengthening AI Regulations And Incentives

The final lever in our original paper focused on the need for governments around the world to develop the vision, skills, and capacities required to effectively regulate AI. Though industry norms and consumer demand play a major role in advancing trustworthy AI, we won't get far without policies to incentivize more responsible practices, and legal mechanisms to hold companies accountable.

Positive Progress→

Our fellow advocates and researchers have been clamoring for action on AI-related regulation for years, but the generative AI boom has made these calls impossible to ignore. Widespread consumer awareness of AI has put more pressure on lawmakers to get up to speed. There's more momentum than ever behind global policy efforts to develop and implement effective and thoughtful AI regulations.

The EU is moving more quickly than some other regions. Lawmakers there have agreed in principle on the EU AI Act, a first-of-its-kind piece of legislation originally proposed in April 2021. The framework takes a predominantly risk-based approach to regulating AI, with separate rules for the most powerful general-purpose AI models. Though there are some limitations to this approach, the EU AI Act is set to become the most comprehensive AI law in the world, and will have significant impacts on global AI governance efforts. The law will complement other recent European technology laws including the Digital Services Act (DSA) and the Digital Markets Act (DMA).

Since its inception, our policy and advocacy teams played a key role in the development of the EU AI Act. Working with our allies, we successfully pushed for more transparency, binding rules for foundation models, and targeted due diligence obligations along the AI value chain. However, the work isn't done yet. We'll continue to advocate to make it a success until the law is fully implemented.

In the U.S., where many of the biggest names in AI are headquartered, the regulatory discussion is starting to pick up. The Biden Administration is looking to move from voluntary AI safety commitments toward concrete rules. In October 2023, President Biden released his Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence with directives to enhance safety, privacy, equity, and competition — providing a welcome move to ensure AI development comes with sufficient regulatory guardrails. In November 2023, Vice President Harris announced draft policy guidance to mitigate risks when the federal government uses AI, putting the federal government's purchasing power behind shaping better AI norms in industry.

Strengthening AI Regulations and Incentives

Though the U.S. has failed to bring comprehensive privacy legislation to fruition, lawmakers are eager to get AI regulation right. In addition to the voluntary commitments AI companies made at the White House in summer 2023, Senate Majority Leader Chuck Schumer held a series of closed-door AI Insight Forums for lawmakers, featuring tech CEOs, researchers, and civil rights leaders. It's crucial that a broad diversity of voices are heard in these forums and it's encouraging that the policy community is seeking out AI expertise, including from Mozilla and its fellows, to develop better legislation.

In Mozilla's written statement for the forum, we highlighted the need for AI policy to center privacy, openness, and transparency as the backbone of responsible regulation. We urged policymakers to look beyond a binary notion of open versus closed AI. A balanced environment where both ecosystems can flourish will fuel innovation, ensure competitiveness, and protect people's rights and safety while mitigating potential risks. We also emphasized the need for lawmakers to champion privacy in AI technologies by default, with comprehensive privacy legislation like the proposed American Data Privacy and Protection Act at the forefront.

Our leaders are also offering their expertise in the U.K. Mozilla.ai gave oral evidence to the U.K.'s House of Lords Communications and Digital Committee as part of their inquiry into LLMs. In our remarks, we emphasized the role that open approaches to AI can play in innovation, the market failures preventing smaller players from accessing computing resources like GPUs, and the need for more government investment in AI infrastructure to promote competition. We also discussed the importance of digital education for enterprises, schools, and civil services on what these models are capable of, and how to deploy them safely in various contexts.

“

Any regulatory framework should ensure that the AI market remains open to competition and innovation from companies challenging the industry behemoths.

”

Work to be Done →

Regulation is needed to make AI more trustworthy and mitigate the risks of the technology. At the same time, regulators need to be mindful of the impact such rules will have on competition and openness in AI. Any regulatory framework should ensure that the AI market remains open to competition and innovation from companies challenging the industry behemoths. To do so, it must safeguard openness and open source.

Openness and transparency are key if we want the benefits of AI to reach the majority of humanity, rather than seeing them applied only to use cases where profit is the primary motivator. Recent global policy discussions on openness have lacked nuance — partly due to outsized influence from big tech incumbents trying to shape regulatory discussions to their benefit. Makers of proprietary models have cited hypothetical catastrophic threats as the most important issues for lawmakers to focus on, neglecting existing AI harms like bias and discrimination. In October 2023, we and more than 1,800 signatories pushed back on this dynamic in our Joint Statement on AI Safety and Openness:

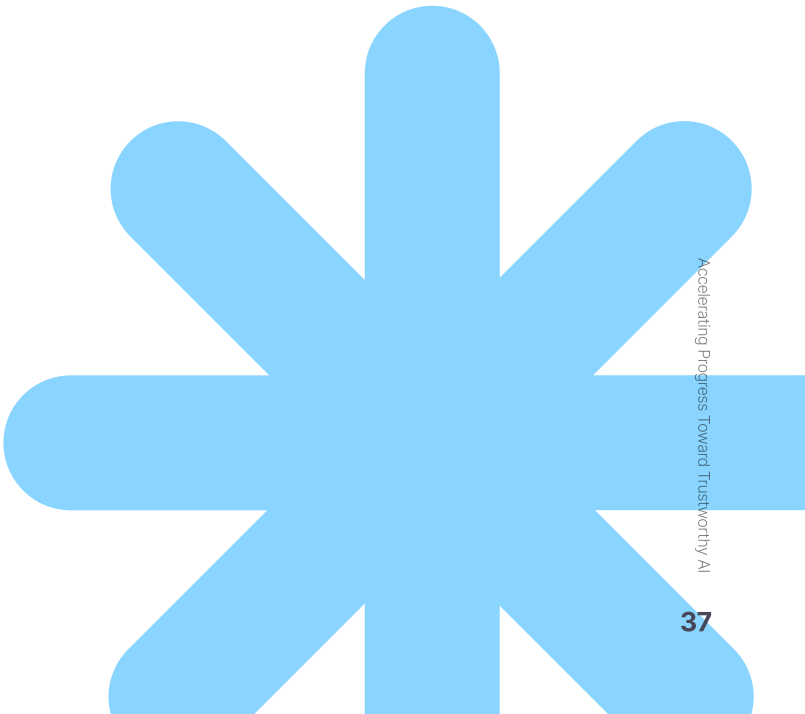
> Yes, openly available models come with risks and vulnerabilities — AI models can be abused by malicious actors or deployed by ill-equipped developers. However, we have seen time and time again that the same holds true for proprietary technologies — and that increasing public access and scrutiny makes technology safer, not more dangerous. The idea that tight and proprietary control of foundational AI models is the only path to protecting us from society-scale harm is naive at best, dangerous at worst.

In October 2023, we and more than **1,800** signatories called for a deeper embrace of open source and open science in AI.

There are government-led processes that can help us sort through some of these bigger questions on open source governance and regulation. As directed in the Biden Administration's executive order, the U.S. Department of Commerce's National Telecommunications and Information Administration (NTIA) is reviewing both the risks and benefits of openly available LLM model weights, inviting public comments to inform potential regulatory approaches. Mozilla intends to submit a response to the associated request for comment to inform NTIA's approach to this issue.

When paired with consumer protections and strong rules to prevent anti-competitive practices, openness spurs innovation and accelerates competition by providing common resources for the ecosystem at large. Competition spurs investments, new jobs, and better choices for companies and consumers. To extend the benefits of the AI boom beyond big tech, lawmakers must prioritize enforcing and strengthening existing competition rules to better meet the challenges of today.

Additionally, while lawmakers educate themselves on the intricacies of the AI landscape, bad actors are poised to weaponize generative AI tools to sow disinformation and political unrest — a harm that is happening right now. With more than 40 national elections scheduled for 2024, policymakers must move more quickly to address this year's threats. We must use the opportunity to study and engage with AI's impacts on global politics, and strengthen our systems for the next set of elections. To that end, Mozilla is highlighting the work of researchers around the world who are uncovering inequities in how platforms approach global elections. We're spotlighting the 'copy-and-paste' policy approach platforms tend to take to global elections, particularly for countries in the Global Majority, and showing the devastating impact such decisions can have on a country's information ecosystem, especially where democratic institutions are relatively fragile.



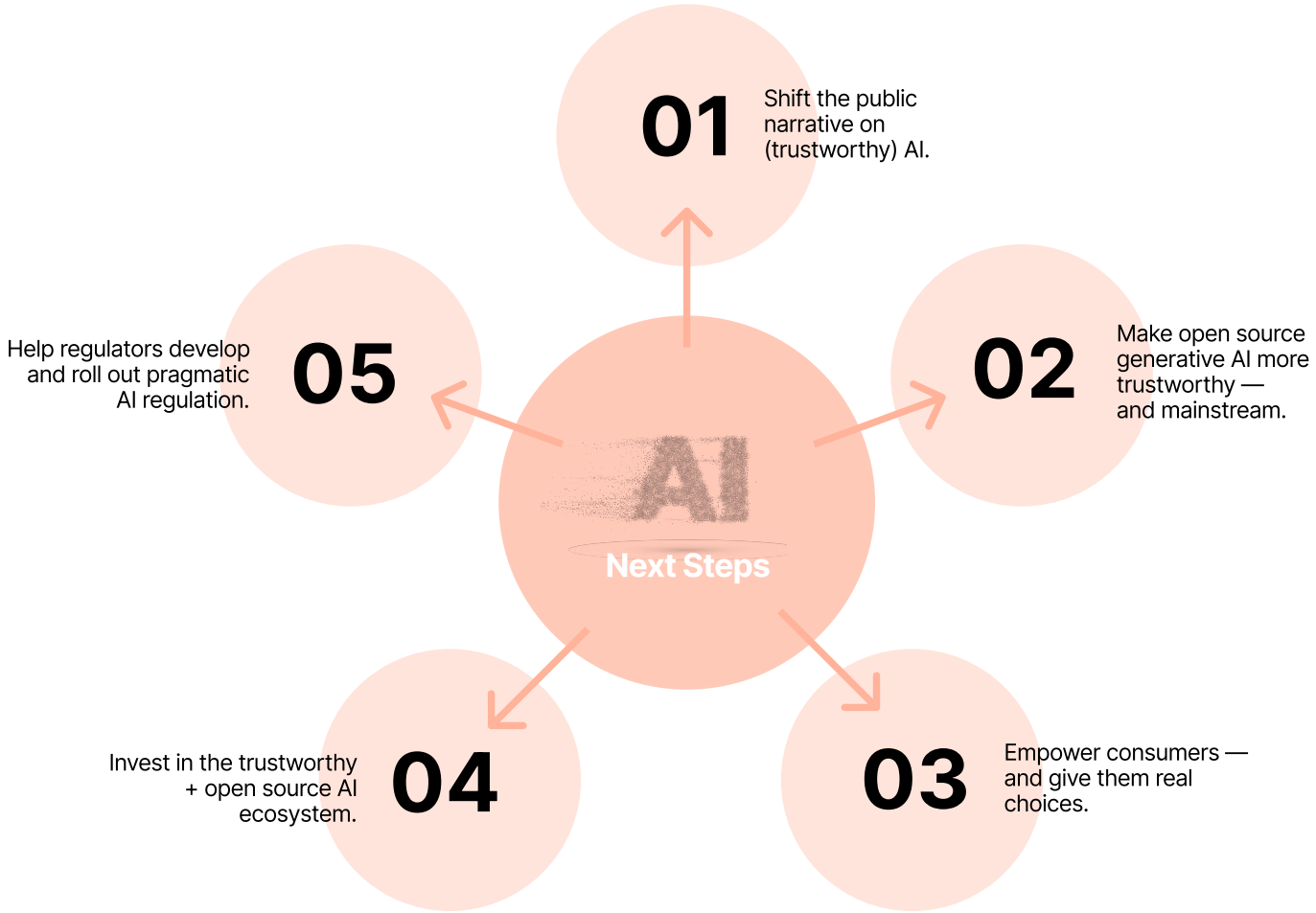
07

The Path Forward For Trustworthy AI

The AI landscape is moving more quickly than ever, and the trustworthy AI ecosystem is growing alongside it. To build on the positive momentum we've seen in the last three years, we must take targeted action across each of our four key levers: industry norms, new tech and products, consumer demand and regulations and incentives.

Next Steps for Mozilla→

We will continue our work to earn users' trust in AI across the entire Mozilla project, focusing on openness as our guiding principle. By scaling the potential of open source approaches and advocating for fair, open markets, we can realize our vision of a trustworthy AI landscape with agency and accountability at the center.



Here's What's Next on Our To-Do List:→

01

Shift the public narrative on (trustworthy) AI.

We have an opportunity to unlock huge benefits from AI — and an urgent need to tackle tough questions about social ills and closed markets. Mozilla will work with activists, builders and policymakers to make sure this more nuanced story about AI breaks through. We're working with civil society organizations around the globe to build sustained political power and shift the narrative. We're also spotlighting the voices of responsible builders, technology leaders, and innovators to help them set the AI agenda in media coverage, policy debates, and popular culture.

02

Make open source generative AI more trustworthy — and mainstream.

Open source generative AI is gaining momentum. Our goal is to ensure open source models are trustworthy, safe, helpful and easy to use. Mozilla projects like the Communal AI small language model platform, the Common Voice dataset, and llamafile are local and more accessible models aimed squarely at this goal.

03

Empower consumers — and give them real choices.

In an era when AI is becoming intricately woven into the fabric of our daily lives, we must champion products that prioritize and exemplify the highest standards of privacy, security, and transparency. By drawing on the invaluable insights gained from technologies like Fakespot, we're deepening our dedication to integrating cutting-edge AI capabilities that genuinely empower consumers within our product suite. We're expanding the impact of these efforts through initiatives like *Privacy Not Included, which are instrumental in equipping consumers with the essential knowledge to make enlightened product choices.

04

Invest in the trustworthy + open source AI ecosystem.

No single company or organization can serve as a counterweight to big tech — but a community can. We will continue to expand our investment in startups, open source projects, and nonprofits building trustworthy AI, both through direct investment and through thoughtful grantmaking. Properly resourced, this ecosystem has the potential to challenge the big players and push AI in a better direction.

05

Help regulators develop and roll out pragmatic AI regulation.

The EU AI Act, the U.S. Executive Order on AI, and similar initiatives in other countries show policymakers are serious about trustworthy AI. Mozilla is increasing its resources to help policymakers roll out regulation and policies that are helpful and pragmatic from both policy and operational perspectives. This work includes publishing research on topics like open source AI models, competition, and privacy that policymakers can draw on, convening policymakers to share expertise and experience, and spotlighting positive policy advancements around the globe.

Next Steps for the Trustworthy AI Ecosystem→

We're committed to doing our part, but we can't make trustworthy AI a reality without working together across the entire tech ecosystem. Here's how you can get involved:

01

BUILDERS: Seek out — and contribute to — trustworthy open source AI projects.

Instead of reaching for the most well-known proprietary models, take advantage of advancements in open source LLMs, and learn from curated resources like our AI Guide. As you develop new tools, engage with builders, users, and researchers from a wide range of backgrounds to widen your perspective. Understanding how AI will impact people who don't think like you will make your project or product that much better.

02

CONSUMERS: Be critical — and know there are choices you can make.

Consumers can't control who builds AI, but they can choose more trustworthy products when available, and demand them when they aren't. We know that public pressure can drive change, even within the most powerful companies on Earth. Look past the "cool factor" of new AI tools and read up on the pros and cons before you experiment. Accessible guides like our *Privacy Not Included series offer clear comparisons in plain language to help you make informed decisions about everything from voice assistants to smart home products.

03

POLICYMAKERS: Prioritize openness and accountability in new rules for AI.

Big tech is pushing for LLM licensing regulations, ostensibly as a security measure. But we know from experience that limiting who can access or benefit from new digital technologies does not make us safer. Openness and accountability can be an antidote, and policymakers must shape legislation accordingly.

04

CIVIL SOCIETY ADVOCATES: Look for intersections between AI and issues your communities care about.

Whether an organization focuses on human rights, climate justice, LGBTQ+ rights, or racial justice, AI is relevant. Philanthropy can offset the influence of tech incumbents by supporting the smaller players pioneering AI approaches that uplift society, not just stock prices. Focus grantmaking on projects that center agency, transparency, and accountability in AI, and look to those that connect to the work you're already doing.

05

INVESTORS: Fund companies, organizations and projects focused on Trustworthy AI

It's tempting to put your money on the big industry leaders, but alternative AI business models that put user privacy and well-being first present a massive opportunity. Whether you're a venture capitalist, an institutional investor or a philanthropist, financially supporting the growth of the trustworthy AI ecosystem will help mitigate AI risks while spreading the returns beyond big tech.

“

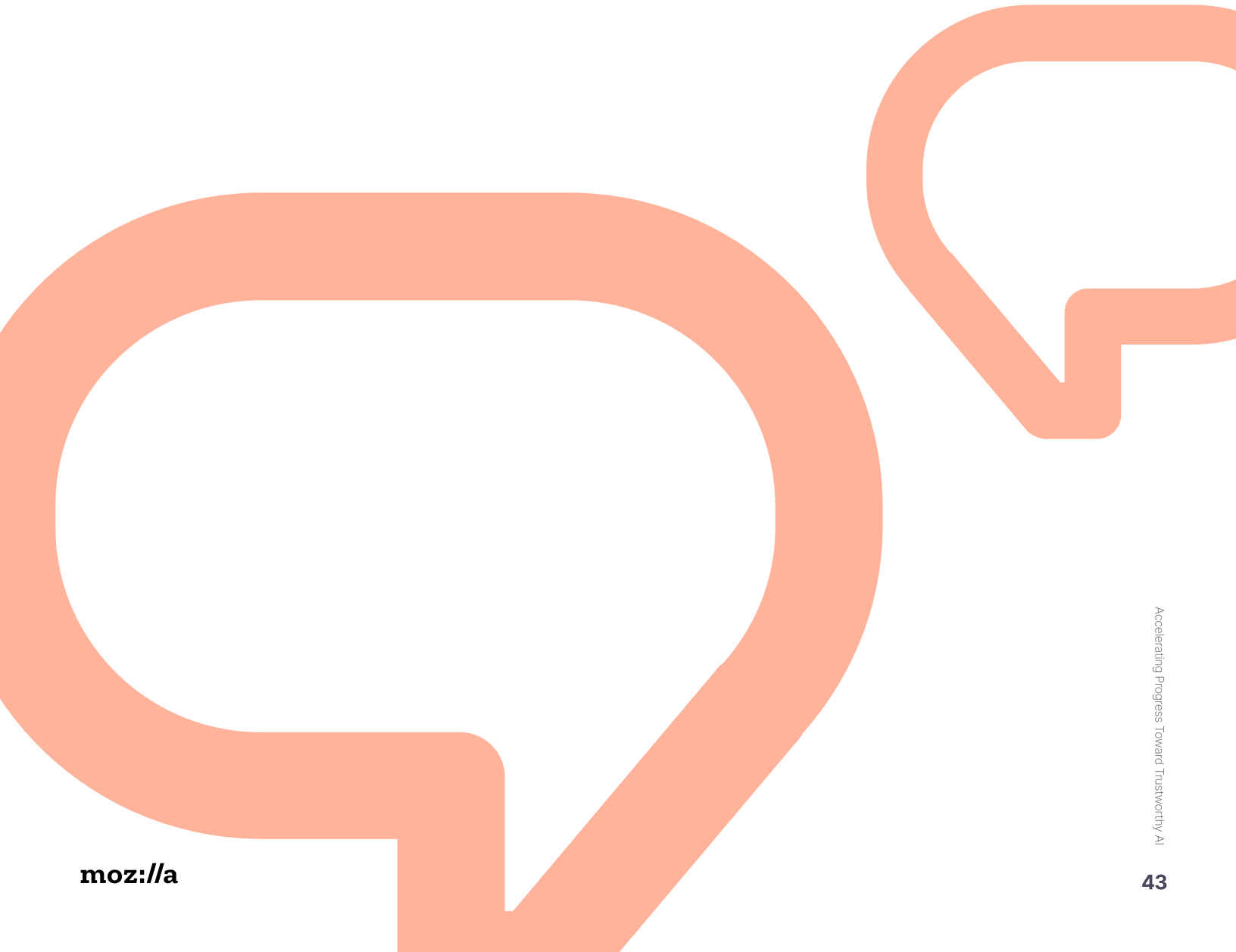
With a movement grounded in openness, agency, and accountability, a more trustworthy AI landscape is within reach.

”

The Path Forward for Trustworthy AI

The AI landscape is moving more quickly than ever, and the trustworthy AI ecosystem is growing alongside it. To build on the positive momentum we've seen in the last three years, we must take targeted action across each of our four key levers: industry norms, new tech and products, consumer demand and regulations and incentives.

Please email us at AIPaper@mozillafoundation.org to provide any input on the report and/or to highlight your favorite examples of AI being used in ways that build trust and improve people's lives.



08

Further Reading

Further Reading

This is the real lesson to take away from the OpenAI debacle, Fast Company (op-ed), December 2023: In an op-ed, Mozilla's Mark Surman explains how OpenAI's November governance battle points to the need for public institutions that prioritize humanity's interests over profit, especially in the AI era, despite the failure of OpenAI's nonprofit model.

When AI doesn't speak your language, Coda, October 2023: Highlights the challenges minority languages face with AI, where better technology could simultaneously support language use and increase surveillance.

AI's Present Matters More Than Its Imagined Future, The Atlantic (op-ed), October 2023: Mozilla Fellow Inioluwa Deborah Raji writes about her experience attending one of Sen. Chuck Schumer's AI Insight Forums, and why present-day harms are more urgent than hypothetical existential AI risks.

How should regulators think about "AI"?, Dr. Emily M. Bender (video), October 2023: Dr. Bender spoke at a virtual roundtable on AI in the workplace convened by Congressman Bobby Scott, breaking down the six different kinds of automation and providing recommendations for AI regulation.

The battle over Open-Source AI, Ben's Bites (newsletter), October 2023: This piece summarizes where well-known AI companies stand on the issue of regulating advanced open source AI software.

Artificial Intelligence: Advancing Innovation Towards the National Interest, Clément Delangue, Hugging Face (written congressional testimony), June 2023: In his testimony, Hugging Face's CEO emphasizes the importance of open AI innovation and the need for mechanisms that ensure AI is safe, transparent, and aligns with national interests.

We tested ChatGPT in Bengali, Kurdish, and Tamil. It failed., Rest of World, September 2023: Rest of World's testing revealed ChatGPT's struggles with many underrepresented languages. The system often makes up words and fails at logic and basic information retrieval, highlighting gaps in AI training data and the need for tailored language support.

The Battle Over Books3 Could Change AI Forever, WIRED, September 2023: This piece details the battle over the Books3 training data set, which was created from a vast collection of copyrighted literary works and is now at the center of disputes between open-access advocates and copyright holders fighting for control and compensation.

LoRA Fine-tuning Efficiently Undoes Safety Training from Llama 2-Chat 70B, LessWrong, October 2023: This study demonstrates how AI models can be easily manipulated to undo safety training, raising concerns about the risks of public model releases.

Removing RLHF Protections in GPT-4 via Fine-Tuning, University of Illinois Urbana-Champaign and Stanford University, November 2023: This study reveals that attackers can remove reinforcement learning with human feedback (RLHF) protections in language models like GPT-4, highlighting the need for enhanced protection against potential misuse.

AI Red-Teaming Is Not a One-Stop Solution to AI Harms, Data & Society, October 2023: This policy brief argues that while AI red-teaming can identify specific technical vulnerabilities, it must be paired with other accountability tools, including algorithmic impact assessments, external audits, and public consultation.

DeepMind reportedly lost a yearslong bid to win more independence from Google, The Verge, May 2021: Google rejected DeepMind's request for greater autonomy and nonprofit status, due to the AI subsidiary's ongoing financial losses and Google's desire to commercialize its AI research.

These fake images reveal how AI amplifies our worst stereotypes, The Washington Post, November 2023: AI image generators like Stable Diffusion and DALL-E continue to perpetuate disturbing stereotypes related to gender and race despite attempts to detoxify their training data, illustrating the urgent issue of inherent bias in AI systems.

OpenAI is getting trolled for its name after refusing to be open about its A.I., Fortune, March 2023: Fortune details the criticism OpenAI has faced for its use of "open" language despite its focus on proprietary, closed source models.

Meta can call Llama 2 open source as much as it likes, but that doesn't mean it is, The Register (op-ed), July 2023: Steven J. Vaughan-Nichols argues that Meta's release of Llama 2 under a "community license" falls short of open-source principles, making the company's use of the term more about marketing than the principles of the open-source community.

Other Resources

- AI Incident Database: Indexing the collective history of harms by AI
- Algorithmic Justice League Harm Collection Tool: Allows users to report AI harms, biases and triumphs
- The Data Provenance Initiative: Audit of large scale datasets

09 Appendix - Additional Mozilla Trustworthy AI Projects

Appendix - Additional Mozilla Trustworthy AI Projects

Changing AI Development Norms

Responsible Computing Challenge: In 2023, Mozilla provided $2.7M to universities in Kenya, India, and the US to add responsible computing to their curricula. The result: thousands of students — the AI builders of tomorrow — wrestling with ethical issues in tech.

Is that Even Legal? guide: Mozilla is educating AI builders on how to develop trustworthy AI systems within existing regulatory frameworks around the world. Our guide offers data governance research and advice for builders in Germany, India, Kenya and the U.S.

Africa Innovation Mradi: This program leverages Mozilla's role as stewards of the open web to promote innovation grounded in the unique needs of users in the African region beginning with East and Southern Africa.

Mozilla Trustworthy AI Fellowships: Long before ChatGPT, Mozilla Trustworthy AI Fellows were studying AI's flaws, limits, and potential. Since 2019, more than 50 Fellows have explored the impacts of AI on society.

Mozilla Festival (MozFest): Too often, the most pressing decisions about AI get made in silos. Mozfest is Mozilla's antidote to this problem. The event convenes and connects thousands of activists, engineers, philanthropists, and policymakers from around the world to build and envision more trustworthy AI.

Lelapa AI: Mozilla Ventures invested in Lelapa AI, a South African-based that just launched its first product: Vulavula, a new AI tool that converts voice to text and detects names of people and places in written text in four South African languages. Lelapa's CEO, Pelonomi Moila, was recently named to the TIME100 in AI.

Building New Tech and Products

Responsible AI Challenge: In May 2023, Mozilla hosted an event with 175 attendees, 7 workshops, and 3 keynote speakers to explore how our Trustworthy AI principles could be used to provide a playbook for builders. The event awarded $100K in prizes to challenge winners who pitched Responsible AI projects.

Mozilla Internet Ecosystem (MIECO): MIECO funds innovators building a healthier internet experience. Supported projects include llamafile, which makes open source large language models much more accessible to both developers and end users.

Mozilla AI Guide: A community-driven resource where developers can come together to pioneer and drive generative AI innovations.

Mozilla Common Voice: The world's largest multilingual, open-source dataset, Common Voice is used by researchers, academics, and developers around the world to train voice-enabled technology and ultimately make it more inclusive and accessible.

Mozilla Technology Fund (MTF): Since 2022, MTF has supported open source projects that explore how AI impacts issues ranging from bias to climate change. One notable project, Countering Tenant Screening, exposes bias and discrimination within the AI-powered screening services used by landlords.

Mozilla Data Futures Lab (DFL): The race to collect data to build and test AI Models has raised new legal and ethical questions about the source and ownership of data. Mozilla's Data Futures Lab incubates products and platforms radically redesigning what trustworthy data stewardship looks like.

Mozilla.ai: Mozilla has been a key contributor to the NeurIPS 2023 Large Language Model Efficiency Challenge, the Conference on Knowledge Discovery and Data Mining (KDD)'s workshop on the evaluation of recommender systems, and research on new ways to efficiently perform few-shot classification on top of closed models like chatGPT.

Themis AI: Mozilla Ventures invested in Themis AI, a Cambridge, MA-based company that spun out of MIT's CSAIL. Themis tackles bias and uncertainty in AI models and has developed a tool, CAPSA, that can automatically estimate uncertainty for any ML model.

Appendix - Additional Mozilla Trustworthy AI Projects

Raising Consumer Awareness

*Privacy Not Included: *PNI guides expose the realities and risks of connected devices. The 2023 guide focused on cars and generated unprecedented attention: citing our research, US Senator Ed Markey wrote to 14 car companies in the US demanding information about their collection, use and sales of personal data.

IRL Podcast: The latest season of IRL showcased global trustworthy AI innovators from civil society, industry, and policy — connecting the issues they care about to AI. The goal was to remind IRL's growing audience that there's no taking the human out of the algorithm. There have been over 100,000 downloads of IRL since Season 7 launched in October 2023.

Philanthropic Advocacy: In 2023, Mozilla participated in a collaboration with Vice President Kamala Harris' office on philanthropy's role in AI alongside other leading foundations. We also published a set of AI Funding Principles that draw on our 4+ years of funding trustworthy AI.

Open Source Research and Investigation Team: The OSRI team uses crowdsourced data to make opaque and influential AI systems more transparent and accountable. The team has produced several original research reports into YouTube's recommendation algorithm.

Mozilla Innovation Week: In December 2023, Mozilla shared a behind-the-scenes view of some of our AI-driven explorations–including Solo, MemoryCache, AI Guide, llamafile–broadcasting on our AI Discord and the Mozilla Developer YouTube channel. The goal was to share transparently what we're working on and what we hope to accomplish.

MemoryCache: This Mozilla Innovation Project is an early exploration project that augments an on-device, personal model with local files saved from the browser to reflect a more personalized and tailored experience through the lens of privacy and agency.

Strengthening AI Regulations and Incentives

Joint Statement on AI Safety and Openness: Signed by over 1,800 scientists, policymakers, engineers, activists, entrepreneurs, educators and journalists, this open letter calls on global lawmakers to embrace openness, transparency, and broad access to mitigate harms from AI systems.

C2PA Standard: Mozilla Ventures' portfolio company Truepic — a key member of the C2PA — advocates for an open technical standard for content verification, including for generative AI. The C2PA standard will allow publishers, creators and consumers to have the ability to trace the origin of AI-generated content.

EU and US Advocacy Campaigns: Mozilla engaged extensively on the EU's AI Act and with policymakers in the United States around AI risk management and accountability.