moz://a

Creating Trustworthy AI

a Mozilla white paper on challenges and opportunities in the AI era

December 2020

Draft v1.0

Established in 2003, guided by the Mozilla Manifesto, the Mozilla Foundation believes the internet is a global public resource that must remain open and accessible to all. The Mozilla Foundation is a not-for-profit organization that exists to support and collectively lead the open source Mozilla project. It views its work as part of a global movement for a digital environment that aims at putting people in charge of their own data and that makes the internet a more democratic place by mobilizing a critical mass of conscious internet users.
Many staff, fellows, and allies of Mozilla generously contributed data and ideas alongside countless readers who participated.

The report was written by Becca Ricks and Mark Surman.

Contributing authors included: Abigail Cabunoc Mayes; Ashley Boyd; Brandi Geurkink; David Zeber; Frederike Kaltheuner; Ilana Segall; J.Bob Alotta; Jane Polak Scowcroft; Jess Stillerman; Jofish Kaye; Kevin Zawacki; Marshall Erwin; Martin Lopatka; Mathias Vermeulen; Muriel Rovira Esteva; Owen Bennett; Rebecca Weiss; Richard Whitt; Sarah Watson; and Solana Larsen.

First published in 2020 © Mozilla Foundation

Table of Contents

I. README

This white paper unpacks Mozilla's theory of change for pursuing more trustworthy artificial intelligence (AI). In this README section, we provide context and key definitions for understanding the paper.

Definitions

We have chosen to use the term AI because it is a term that resonates with a broad audience, is used extensively by industry and policymakers, and is currently at the center of critical debate about the future of technology. However, we acknowledge that the term has come to represent a broad range of fuzzy, abstract ideas.[1] Mozilla's definition of AI includes everything from algorithms and automation to complex, responsive machine learning systems and the social actors involved in maintaining those systems.

Mozilla is working towards what we call **trustworthy AI**, a term used by the European High Level Expert Group on AI.[2] Mozilla defines trustworthy AI as AI that is demonstrably worthy of trust, tech that considers accountability, agency, and individual and collective well-being.

Mozilla's theory of change is a detailed map for arriving at more trustworthy AI. We developed our theory of change over a one-year period. During this timeframe, Mozilla consulted with scores of AI domain experts from industry, civil society, academia, and the public sphere. We conducted a thorough literature review. We also learned by doing, running advocacy campaigns that scrutinized AI used by Facebook[3] and YouTube;[4] funding art projects that illuminated AI's impact on society;[5] and publishing relevant research in our Internet Health Report.[6]

Mozilla's theory of change focuses on AI in **consumer technology**: general purpose internet products and services aimed at a wide audience.[7] This includes products and services from social media platforms, search engines, and ride sharing apps, to smart home devices and wearables, to e-commerce, algorithmic lending, and hiring platforms.

Limitations

We acknowledge that AI is used in ways that are harmful outside of the consumer tech space: surveillance by governments, facial recognition by law enforcement, and automated weapons by militaries, for instance. Although this is not typically our focus, we care deeply about these issues and support other organizations that are pushing for greater public accountability and oversight of this tech. New technologies are often normalized in consumer technologies before they are adopted in more high-stakes environments, such as governments. In some cases,

companies work hand in hand with governments, blurring the line between public and private spaces and requiring us to explore new avenues for advocacy. By focusing our attention on consumer-facing technologies, we believe we can impact how they are deployed in public contexts.

Civil society organizations such as EFF, AI Now Institute, the ACLU, and Data & Society are exploring how AI used by governments or law enforcement can threaten civil liberties. Journalists and human rights organizations like Amnesty International and Human Rights Watch act as watchdogs to check government abuses of AI, such as the use of autonomous weapons, military drones, or computational propaganda. Mozilla's focus on consumer applications of technology — paired with our technical expertise and history in the tech space — allows us to engage with a slightly different set of questions. In any future explorations of the trustworthy AI space, we will continue to describe how we believe Mozilla's work fits into these critical conversations.

Another limitation: Many examples are drawn from EU and US contexts, especially in discussions of what effective regulatory regimes might look like. We invite critique and examples of positive interventions happening outside of these regions as we continue to build a global, diverse movement.

About Mozilla

The 'trustworthy AI' activities outlined in this document are primarily a part of the movement activities housed at the Mozilla Foundation — efforts to work with allies around the world to build momentum for a healthier digital world. These include: thought leadership efforts like the Internet Health Report and the annual Mozilla Festival, fellowships and awards for technologists, policymakers, researchers and artists, and advocacy to mobilize public awareness and demand for more responsible tech products. Mozilla's roots are as a collaborative, community driven organization. We are constantly looking for allies and collaborators to work with on our trustworthy AI efforts.

II. Executive Summary

If we want a healthy internet and a healthy digital society, we need to ensure that our technologies are trustworthy. Since 2019, Mozilla Foundation has focused a significant portion of its internet health movement-building programs on AI. Building on our existing work, this white paper provides an analysis of the current AI landscape and offers up potential solutions for exploration and collaboration.

What's at stake

AI has immense potential to improve our quality of life. But integrating AI into the platforms and products we use everyday can equally compromise our security, safety, and privacy. Our research finds that the way AI is currently developed poses distinct challenges to human well-being. Unless critical steps are taken to make these systems more trustworthy, AI runs the risk of deepening existing inequalities. Key challenges include:

- **Monopoly and centralization:** Only a handful of tech giants have the resources to build AI, stifling innovation and competition.
- **Data privacy and governance:** AI is often developed through the invasive collecting, storing, and sharing of people's data.
- **Bias and discrimination:** AI relies on computational models, data, and frameworks that reflect existing bias, often resulting in biased or discriminatory outcomes, with outsized impact on marginalized communities.
- **Accountability and transparency:** Many companies don't provide transparency into how their AI systems work, impairing mechanisms for accountability.
- **Industry norms:** Because companies build and deploy rapidly, AI systems are embedded with values and assumptions that are not questioned in the product development lifecycle.
- **Exploitation of workers and the environment:** Vast amounts of computing power and human labor are used to build AI, and yet these systems remain largely invisible and are regularly exploited. The tech workers who perform the invisible maintenance of AI are particularly vulnerable to exploitation and overwork. The climate crisis is being accelerated by AI, which intensifies energy consumption and speeds up the extraction of natural resources.
- **Safety and security:** Bad actors may be able to carry out increasingly sophisticated attacks by exploiting AI systems.

Several guiding principles for AI emerged in this research, including agency, accountability, privacy, fairness, and safety. Based on this analysis, Mozilla developed a theory of change for

supporting more trustworthy AI. This theory describes the solutions and changes we believe should be explored.

Making trustworthy AI a reality

While these challenges are daunting, we can imagine a world where AI is more trustworthy: AI-driven products and services are designed with human agency and accountability from the beginning. In order to make this shift, we believe industry, civil society, and governments need to work together to make four things happen:

1. A shift in industry norms

Many of the people building AI are seeking new ways to be responsible and accountable when developing the products and services we use everyday. We need to encourage more builders to take this approach — and ensure they have the resources and support they need at every stage in the product research, development, and deployment pipeline. We'll know we are making progress when:

1.1. Best practices emerge in key areas of trustworthy AI, driving changes to industry norms.
1.2. The people building AI are trained to think more critically about their work and they are in high demand in the industry.
1.3. Diverse stakeholders are meaningfully involved in designing and building AI.
1.4. There is increased investment in trustworthy AI products and services.

There are a number of ways that Mozilla is already working on these issues. We're supporting the development of undergraduate curricula on ethics in tech with computer science professors at 17 universities across the US. We're actively looking for partners to scale this work in Europe and Africa, and seeking ways to work with a broader set of AI practitioners in the industry.

2. New tech and products are built

To move toward trustworthy AI, we will need to see everyday internet products and services come to market that have features like stronger privacy, meaningful transparency, and better user controls. In order to get there, we need to build new trustworthy AI tools and technologies and create new business models and incentives. We'll know we are making progress when:

2.1. New technologies and data governance models are developed to serve as building blocks for more trustworthy AI.
2.2. Transparency is a feature of many AI-powered products and services.
2.3. Entrepreneurs and investors support alternative business models.
2.4. Artists and journalists help people critique and imagine trustworthy AI.

As a first step towards action in this area, Mozilla will invest significantly in the development of new approaches to data governance. This includes an initiative to network and fund people around the world who are building working product and service prototypes using collective data governance models like data trusts and data co-ops. It also includes our own efforts to create useful AI building blocks that can be used and improved by anyone, starting with our own open source text-to-speech efforts such as DeepSpeech[8] and Common Voice data commons.[9]

3. Consumer demand rises

People can play a critical role in pressuring companies that make everyday products like search engines, banking algorithms, social networks, and e-commerce sites to develop their AI differently. We'll know we are making progress when:

3.1. Trustworthy AI products emerge to serve new markets and demographics.
3.2. Consumers are empowered to think more critically about which products and services they use.
3.3. Citizens pressure and hold companies accountable for their AI.
3.4. Civil society groups are addressing AI in their work.

Mobilizing consumers is an area where Mozilla believes that it can make a significant difference. This includes providing people with information they can use everyday to question and assess tech products, as we have done with our annual *Privacy Not Included Guide.[10] It also includes organizing people who want to push companies to change their products and services, building on campaigns we've run around Facebook, YouTube, Amazon, Venmo, Zoom, and others over recent years. These awareness and pressure campaigns aim to meet people where they are as internet users and citizens, giving them even-handed, technically accurate advice. Our hope is that this kind of input will encourage tech companies to develop products that empower and respect people, building new levels of trust.

4. Effective regulations and incentives are created

Consumer demand alone will not shift market incentives significantly enough to produce tech that fully respects the needs of individuals and society. New laws may need to be created and existing laws enforced to make the AI ecosystem more trustworthy. To improve the trustworthy AI landscape, we will need policymakers to adopt a clear, socially and technically grounded vision for regulating and governing AI. We'll know we are making progress when:

4.1. Governments develop the vision and capacity to effectively regulate AI.
4.2. There is wider enforcement of existing laws like the GDPR.
4.3. Regulators have access to the data they need to scrutinize AI.

4.4. Governments develop programs to invest in and procure trustworthy AI.

Mozilla has a long history of working with governments to come up with pragmatic, technically informed policy approaches on issues ranging from net neutrality to data protection. We also work with organizations interested in advancing healthy internet policy through fellowships and collaborative campaigns. We will continue to develop this approach around the issues described in this paper, such as encouraging major platforms to open up their data and documentation to researchers and governments studying how large-scale AI is impacting society. Europe and Africa will be our priority regions for this work.

Working like a movement

Developing a trustworthy AI ecosystem will require a major shift in the norms that underpin our current computing environment and society. The changes we want to see are ambitious, but they are possible. We saw it happen 15 years ago as the world shifted from a single desktop computing platform to the open platform that is the web. There are signs that it is already starting to happen again. Online privacy has evolved from a niche issue to one routinely in the news. Landmark data protection legislation has passed in Europe, California, and elsewhere around the world, and people are increasingly demanding that companies treat them — and their data — with more care and respect. All of these trends bode well for the kind of shift that we believe needs to happen.

The best way to make this happen is to work like a movement: collaborating with citizens, companies, technologists, governments, and organizations around the world. With a focused, movement-based approach, we can make trustworthy AI a reality.

III. Introduction

1. What if AI worked differently?

The prevailing computing environment of any particular era shapes what people think is possible and, in turn, what technologies we build.

When the personal computer became mainstream in the 1980s, we saw the invention of spreadsheets and word processors and, with these inventions, the transformation of the workplace. When the web became widely accessible in the late 1990s and early 2000s, the browsers and, eventually, cloud-based apps became ubiquitous, leading to huge shifts in how we entertain ourselves, collaborate with colleagues, and do business.

Today's computing environment is increasingly shaped by AI and the data that powers it. From recommendation engines to smart email filters to predictive text, AI-powered systems have become ubiquitous in our modern society. The norms around how AI is developed transform what kinds of tools, platforms, and experiences we end up building. As so much of our lives become digital, we will continue to see these norms shape our everyday lives.

For example, one current norm is that companies develop products and services that collect as much data about people as possible, and then use sophisticated models to analyze that data and provide personalized experiences. The results of this norm can sometimes be delightful: Spotify suggests songs we like, and Gmail's autocomplete feature finishes our sentences. But the results can also be harmful: There is evidence that video recommendation engines like YouTube, which optimize for user engagement, profit by introducing people to increasingly extreme viewpoints. In addition, targeted advertisements on Facebook have been shown to manipulate people and exclude vulnerable communities.

Another computing norm is that companies with access to the most data have a competitive advantage in the AI landscape, incentivizing further data collection. Big tech companies have an outsized advantage over both smaller competitors and the people who use tech. Smaller companies find it almost impossible to access enough data to compete on the personalization or recommendation front, and people are often locked into one platform.

As these two examples illustrate, our current paradigms for building technology limit what we think is possible. What if we radically adjusted these norms in AI development? What would it look like if people had greater control over the data collected about them? What kinds of processes and tools in the AI development pipeline will lead to greater accountability?

If our current computing environment is not working, then we must invent a new one. If people feel like they're not in control of their own data, we can incentivize companies to build technologies that give people more agency. By changing the rules around how data is collected and stored, we can invite smaller players to participate. By imagining new processes for how technology is developed, we can shape the platforms, tools, and products that strengthen collective well-being.

Changes like these are necessary if we want AI that strengthens – rather than harms – society and communities.

2. The current AI landscape

Recently, AI has become an area of focus for people, industry, and regulators. The benefits of such technologies are clear: The application of AI to healthcare, law, education, and everyday tasks like email could completely transform the efficiency and accuracy of our current decision-making systems. At the same time, the rapid development of AI by large tech companies has outpaced regulation, sometimes leading to harmful or exploitative outcomes.

People

As our technologies are becoming more complex, many people feel increasingly powerless because companies do not provide them with the information they need to make educated choices, nor do people have a way to know whether a company took steps to prevent harassment, correct for bias, or assess risk before deploying AI. According to a recent poll by Consumers International and the Internet Society, "75% of people distrust the way data is shared" and "63% of people find connected devices 'creepy.'"[11] Meanwhile, recent research by the Knight Foundation reveals that Americans believe big internet companies "create more problems than they solve."[12] And an Amnesty International survey shows that a majority of people on five continents worry about big tech abusing their data.[13]

Companies widen this knowledge gap by actively presenting their technology as too complicated for people to understand, through labyrinthine privacy policies and other mechanisms that obscure what is actually happening. When people do not understand what's happening to their data, they may not feel empowered to advocate for clarity or control. In addition, companies might not be incentivized to address harms that are experienced by a minority of their users or experienced by communities that do not even use the product. We need to work towards a more inclusive understanding of what constitutes harm and exploitation in order to address the impact of AI on people and society.

Industry

Current incentives in the tech industry have resulted in business models that rely on unfettered access to data. Within the surveillance economy, data about people's behavior — from what we search for, to where we travel, to what we buy — can be collected and exploited with few restrictions.

At the same time, the industry is dominated by a handful of tech giants who wield immense market and political power. Companies like Amazon, Google, Apple, Microsoft, and Facebook possess vast troves of data about how people interact with each other and the internet. Our data-driven economy rewards these business models, cementing existing tech oligarchies. As AI becomes more pervasive, we are likely to see tech giants continue to stockpile data and reap the benefits.

On the other hand, the people who are building AI — from developers and data scientists to designers and policy analysts — have expressed they want to create more responsible technology. There is growing evidence that AI engineers and students no longer want to work for companies like Facebook or Palantir, whose questionable decisions around data exploitation or military contracts have undermined their public reputation.[14] But even if the people building AI want to do so ethically, they often lack the tools or mandate to do so.

Companies that develop AI have responded to these challenges by creating ethical AI guidelines. While this is a positive development, these guidelines are frequently non-binding and ineffective. For example, Google has committed publicly to avoiding bias in its algorithms, but in the past has used unethical data collection processes in order to diversify its datasets.[15] Further, AI ethics initiatives tend to operate within a neoliberal framework that emphasizes individual actions rather than system-wide change. Some companies have appeared to embrace ethical AI — as long as it remains the responsibility of individuals and doesn't interfere with business.

Regulators

Current norms around AI development have outpaced regulations in many countries, resulting in an environment where technologies and ideas are tested on and deployed to millions of people without proper oversight or transparency. Many countries still lack effective data protection and consumer protection laws, regulation that will be necessary to effectively regulate AI.

There has been small progress: In 2019, 42 countries came together to endorse a global governance framework on AI.[16] Signed by both OECD members and non-members, the accord represents a first step as governments begin to work through the ethical challenges that arise from developing AI.[17] However, we are far from reaching a global consensus on what an effective regulatory framework for AI should look like.

In countries like the US, large tech companies like Google and Microsoft have so far capitalized on self-regulation. Because regulations in the US tend to be narrow in the way they are drafted and interpreted, some of the worst cases of consumer data exploitation often skirt direct regulation.

In regions such as China, Brazil, the UK, and the EU, regulators have taken a more hands-on approach to establishing AI guidelines. In the EU, privacy regulations like the GDPR have transformed the way many tech companies approach consumer data. More recently, the European Commission's 2020 white paper[18] laid a foundation for regulating AI, which includes legal requirements for ensuring datasets are representative and unbiased; keeping detailed documentation of how AI was developed; and requiring more government oversight for AI. As they begin drafting laws in 2020, EU regulators aim to take a risk-based approach to governing AI, focusing on applications in "high risk" areas like healthcare and immigration.

However, the EU fails to classify AI in many technologies as high risk. The use of AI in these products and services creates significant collective risks related to bias, misinformation, and corporate surveillance. The kind of risks posed by these technologies are fundamentally different, however, and might be better addressed by enforcing existing laws, and through dedicated and targeted regulatory interventions.

Similar to the EU's GDPR, Brazil's General Data Protection Law (LGPD), which came into effect in August 2020, affords consumers greater rights over the data companies process about them.[19] Similarly, in India a sweeping Personal Data Protection Bill forces internet companies to seek permission for the use of people's personal data. However, the bill places few restrictions on the government's use of sensitive data as part of the Aadhaar national ID system.[20]

This landscape of dis-empowered people, industry centralization, and floundering regulators can appear bleak. But we have faced similar challenges in the past. Two decades ago, the decentralized web was almost commandeered by the tech giant Microsoft. This threat spurred people, developers, and regulators to push back and imagine something better – and it worked.

3. Mozilla's approach to trustworthy AI

Mozilla has a rich history of reimagining computing norms to favor openness and innovation. We first did this in the early 2000s by championing an openness in an era where the web was on the brink of being monopolized by a single company, Microsoft, which had gone from being a minor player in the browser market to near-dominance. The market dominance of Internet Explorer threatened to lock in users, stamp out competitors, and stifle innovation online.

In the face of Microsoft's monopolization of the browser market, a loose coalition of open source activists, software developers, and web enthusiasts came together to build standards-based browsers and web servers that would eventually wrest power away from the

tech giant. Mozilla was an early and active member of this movement. We focused resources, coordinated code, and ultimately released Firefox as part of this movement. Around the same time, the US Department of Justice's antitrust case against Microsoft demonstrated how regulators can help keep the technology industry competitive and healthy.

The result was a fundamental shift in the computing environment of the time. A renewed interest in web standards like HTML and JavaScript made true cross-platform applications the norm, replacing the dominant paradigm of end user apps that only worked on Windows. This fostered an open environment that allowed new cross platform products and services — including Facebook and Gmail — to enter the field. The internet we know now would not exist if the constrained environment of Windows and Internet Explorer 6 had become the status quo.

Today, we are at a similar inflection point. As in the early 2000s, many of our current problems are caused by a limited playing field. There are bright spots: A growing number of software developers, activists, academics, designers, and technologists are asking critical questions about how current norms around AI and data are centralizing power, stifling innovation, and eliminating user agency. But these efforts desperately need more fuel.

In this paper, we provide Mozilla's perspective on how we might do just this. Our work began in earnest in 2019, when members of the Mozilla community began asking questions like: What can Mozilla do to shift norms around AI? Who else is tackling this problem? And, how can we help them? We emerged from the exploration process with big-picture learnings. For instance, while many of the challenges with AI are individual, large scale AI also presents major collective risks. We also emerged with granular learnings. For instance, there is progress being made in creating privacy-preserving ways to handle data for machine learning. In addition, governments are hungry to figure out how to fairly and effectively regulate AI, but they lack the internal expertise and independent research needed to do so.

All of these learnings culminated in Mozilla's theory of change — a rough road map for what levers we need to pull in order to achieve trustworthy AI at scale and in a lasting way. Some of these levers exist in the realm of industry: Mozilla can support better education for computer science students or push for greater algorithmic accountability. Some of these levers exist in policy: We can steer more like-minded technologists toward government or advocate for stricter enforcement of privacy laws. Other levers exist in civil society, in the realms of academia, activism, art, and journalism.

All these levers are interconnected, and over the coming years, Mozilla will focus our effort and resources on pulling these levers. However, we know that our own contribution to this work exists within a much larger constellation of actors. Just like we did in the early Firefox era, Mozilla will function as one part of a broader movement: focusing resources, coordinating work, and nurturing a more equitable computing environment.

IV. Challenges with AI

Before answering the question of "what would trustworthy AI look like?" we need to examine the unique challenges that AI presents to the tech industry and society at large.

A significant body of work has emerged on these challenges over the past few years, revealing that AI can exhibit bias and invade privacy. As it is woven into all aspects of our lives, AI has the potential to reinforce existing power hierarchies and societal inequalities. This raises questions about how to responsibly address potential risks for individuals in the design of AI.

More recently, a further body of work is emerging around the collective risks and harms associated with the widespread adoption of AI. Some liken these collective risks and harms of AI to pollution or climate change, since they impact all of us on a massive scale, and can only be addressed by looking at the ecosystem as a whole.

Finally, there are more technical challenges that relate to the current design norms of AI itself. For example, many have noted that AI techniques resist oversight because they lack mechanisms for transparency. Issues like this are often seen as flaws for the industry to tackle as a whole.

In the section that follows, we don't aim to provide a complete picture of all the problems posed by AI. Rather, this analysis represents our own assessment of those challenges AI poses to society within a consumer tech context.

1. Monopoly and centralization

Only a handful of tech giants have the resources to build AI, stifling innovation and competition.

AI is poised to transform how we work, socialize, learn, and interact with one another. At the same time, these transformative technologies are being developed by only a handful of large companies, resulting in a market for AI that isn't truly competitive or innovative. Currently, corporations like Google, Amazon, Apple, IBM, Microsoft, Facebook, Baidu, Alibaba, and Tencent — described as the "Big Nine"[21] — exercise the most power over the AI market.

Companies have a tendency to stockpile data in order to maintain their competitive advantage. Once AI enters the equation, though, it creates an endless cycle: Those companies who dominate the market have greater access to data, which allows them to develop better machine learning models, which encourages users to use the platform and generate more data. Amazon,

for instance, is currently using AI to improve how its AWS cloud computing business runs,[22] further cementing the company's hold over the market.

For "platform monopolies" like Facebook and Google that amass huge troves of data about how people behave online, the competitive advantage is even more pronounced. Facebook, for instance, owns all the data it collects from users on its platform and uses that data to build increasingly complex AI, such as its personalized News Feed feature and ad targeting ecosystem. The companies who dominate the AI space have no incentive to share data back with the public, reinforcing this power asymmetry.

Rapid consolidation of the AI space is likely to continue, as the most dominant tech companies acquire their AI competitors and the data that come with them. For instance, Facebook has acquired former competitors Instagram and WhatsApp. In 2019, Google's $2.1 billion acquisition of Fitbit, the maker of smartwatches and fitness trackers, was widely viewed as a move to expand into the healthcare sector by amassing more health data.[23]

Many of these companies have recently come under scrutiny from lawmakers globally to determine whether or not they are violating antitrust laws. The EU has launched a number of antitrust probes into tech companies, and in 2017, the EU Commission fined Google €2.4 billion for favoring its own search services over its competitors'. In 2019, the US House Judiciary antitrust subcommittee sent letters to Amazon, Apple, Alphabet (the parent company of Google), and Facebook out of concern that such companies hold too much market share.[24]

Regulatory solutions have been proposed, including stricter enforcement of antitrust laws or enacting new oversight laws. Others have suggested nationalizing the "platform monopolies" so that they more fully serve the public interest.[25] Additionally, alternative data governance models like data trusts have been proposed to shift data ownership from platforms back to users.

2. Data governance and privacy

Because AI requires access to large amounts of training data, companies and researchers are incentivized to develop invasive techniques for collecting, storing, and sharing data without obtaining meaningful consent.

In the decades spent developing the online advertising ecosystem, companies have engaged in invasive data collection without meaningful user consent in an effort to amass data and gain a competitive edge, all while skirting accountability. The ubiquity of complex, invasive ad targeting on the web has led many internet users to begrudgingly accept that large tech companies have access to their data.

These privacy concerns intensify with the development of AI. Vast amounts of training data — which may include images, text, video, or audio — are required to teach machine learning

models how to recognize patterns and predict behavior. Copyright laws, privacy rules, and technical hurdles significantly limit what kind of data may be used or purchased by developers. As such, typically developers need to seek out new sources of data in order to train their models.

The current competitive marketplace for machine learning incentivizes companies to collect user data without obtaining meaningful consent and without sufficient privacy considerations. For instance, in 2019 Google suspended its facial recognition research program for the Pixel 4 smartphone after a report revealed that its contractors had been targeting homeless Black people to capture images of their faces through blatant deception.[26]

Even when digital services and platforms do legally obtain user consent to collect data, often it is through default settings, manipulative design, Terms of Service agreements that few people read, or privacy policies written in inaccessible, complex language. Until recently, companies building AI-powered voice assistants like Amazon Alexa and Google Home did not explicitly inform people that their voice interactions may be listened to by human workers to develop the models. Despite changes to these review programs, consent is still couched in vague language. For instance, Amazon Alexa users agree to having their voice recordings reviewed with a toggle that simply says “help improve Amazon services and develop new features.”[27]

As AI continues to drive up the value of people’s data, information asymmetry will continue to increase between users and the companies collecting their data.[28] Some of the most egregious behaviors from companies were made illegal in the EU under the GDPR and would likely be penalized in today’s regulatory environment. However, in countries without strict privacy laws many of these practices may continue unchecked, and even with GDPR limitations in place, companies may continue to collect data without obtaining meaningful consent. It is unclear, for instance, whether individual requests for deletion of personal data filed under the GDPR may apply to models trained on personal information. Questions continue to emerge around what control users truly have over their own data in the current computing environment and what appropriate agency should look like.

3. Bias and discrimination

AI relies on computational models, data, and frameworks that reflect existing bias, often resulting in biased or discriminatory outcomes, with outsized impact on marginalized communities.

Every dataset comes with its own set of biases, and it is impossible to build a fully unbiased AI system. Humans are biased, and every part of the research, collection, structuring, and labeling of data is shaped by human decisions. Bias is the result not just of unbalanced training data, sampling, and data availability, but it is also the product of systemic and methodological choices teams make when they are designing an AI system.

Sometimes the bias exhibited in an AI system is the result of incomplete, unbalanced, or non-representative training data. As computer science researchers Joy Buolamwini and Timnit Gebru have demonstrated,[29] common facial recognition systems routinely misidentify Black faces due to a lack of diversity in their training data. Similarly, scholar Safiya Umoja Noble has written about how searches for the term "professional hairstyles" in Google returned images of white, blonde women, whereas "unprofessional hairstyles" returned images of Black women.[30] In both cases, the technology further entrenched existing racial inequities, marginalizing Black communities and experiences. Due to the outsized impact bias has on marginalized communities, any approach to tackling bias must involve voices and organizations from the racial justice, gender justice, and immigrant justice movements.

Other times, the bias is systemic – the product of the methodological choices made in the design of the AI system. For instance, many ML teams use performance metrics as a benchmark for success in developing and deploying AI systems. If the team decides to set their model's success threshold at 99.99%, then that means that failing to perform correctly for 0.01% of the representative population is the expected behavior. These systems will always exclude or fail some users – *by design*.

Systemic bias is often implicit to the design choices teams make when designing and deploying AI systems. For instance, a system that is trained to be successful for most cases may still end up unintentionally latching onto the "wrong" things in the dataset for a small number of edge cases. This is particularly concerning when such edge cases occur for groups that are already marginalized or oppressed. In 2020, Facebook's automated content moderation system accidentally flagged posts from Nigerian activists protesting the Special Anti-Robbery Squad (SARS), a controversial police agency that activists say routinely carry out extrajudicial killings against young Nigerians, because the acronym "SARS" was linked by Facebook's algorithm to be misinformation about the COVID-19 virus.[31]

Even when steps have been taken to reduce bias in a model, that system can still make decisions that have a discriminatory effect. For instance, Facebook has been criticized for allowing advertisers to discriminate against users belonging to protected groups, like ethnicity and gender, through its targeted advertising platform. However, even when Facebook changed its ad platform to prevent advertisers from selecting attributes like "ethnic affinity" for categories like housing or jobs, it was determined that the platform still enabled discrimination by allowing advertisers to target users through proxy attributes.[32]

Increasingly, computer scientists are now rallying around values like "fairness, accountability, and transparency"[33] and proposing new statistical models for reducing bias. At the same time, we must continue to question the core values the AI system is optimizing for, how the system is designed, or whether such a system should ever be built at all. Any efforts to address bias and discrimination in AI must work with those communities most impacted by such systems.

4. Accountability and transparency

Companies often don't provide transparency into how their AI systems work, impairing legal and technical mechanisms for accountability.

Many platforms develop closed algorithms that rapidly generate, curate, and recommend content. Facebook and Amazon, for instance, curate organic and sponsored content based on what its algorithm predicts we might like to see, share, read, or purchase in order to nudge us towards a desired behavior. This curation of social platforms creates an environment in which ad targeting, filter bubbles, bots, and harmful content thrive, deepening our susceptibility to behavioral manipulation and misleading, polarizing, or inflammatory information. YouTube's recommendation engine is particularly alarming, with evidence that the "autoplay" function pushes viewers towards increasingly inflammatory and conspiratorial extremist content.[34]

Transparency has different use cases for different audiences. For AI developers, transparency means clarifying how technical decisions were made during the design and development of an ML model. Such transparency may only be useful to experts who have the expertise and experience to understand and audit such decisions. Understanding why a model predicted a particular outcome is critical for developers, both to ensure the model is making decisions correctly, and to prevent harmful outcomes. Many computer scientists are actively developing tools to improve the explainability of AI — why a particular prediction was made for a given input. Different definitions of explainability are currently used by developers, and there are no formal evaluation criteria for putting explainability into action.[35]

To end users, transparency could mean conveying the most important points to a broad audience, presenting accessible summaries of what the model is doing. In an ethnography of AI builders,[36] developers said they wanted to establish greater trust with users by showing the ways in which human decisions were made in the development of the system and by building transparency tools people can use.

To watchdogs and policymakers, transparency is only meaningful if tied to clear pathways to accountability. In order to hold AI systems accountable, different stakeholders will need access to different types of information about the system. A social science researcher, for instance, may need access to the targeting criteria of an advertising algorithm in order to audit whether or not the system is discriminatory. A policymaker may need access to documentation about how a content moderation algorithm interacts with humans to make decisions. Transparency efforts are only effective when the preconditions for accountability already exist.[37]

In order to hold companies accountable for how particular AI systems were designed and developed, we will need to continue exploring legal, technical, and institutional mechanisms for accountability.

5. Industry norms

Companies are pressured to build and deploy AI rapidly without pausing to ask critical questions about the human and societal impacts. As a result, AI systems are embedded with values and assumptions that are not questioned in the product development life cycle.

The dominant narrative in tech is to disrupt, "break things," and innovate with increasing speed. This idealism — paired with weak legal limits on what such companies are permitted to do — has allowed for rapid experimentation and deployment of new ideas. But it has also contributed to a culture in which new products are not subjected to critical examination, sufficient testing, or regulatory oversight.

The result is that often AI systems are built under a set of assumptions that have gone unchallenged, and companies optimize for a narrow set of values, such as profitability, engagement, and growth. For instance, YouTube's recommendation algorithm was initially built to optimize for user engagement and not, say, values like user satisfaction and happiness.[38]

This attitude has led to the development of many well-intentioned but problematic technologies that deepen societal inequality. For instance, Uber was founded on the "disruptive" idea of a sharing economy in which its platform would generate new income opportunities. But in reality it relies on the exploitation of freelancers competing with each other for low wages in a hyper-competitive environment governed by algorithms.[39]

A real lack of diversity (professional, cultural, ethnic, gender, socioeconomic, and geographic) contributes to this problem, since the viewpoints offered in decision-making spaces tend to be homogeneous. At companies like Facebook and Google, women make up only 10% and 15%, respectively, of their AI research teams.[40] Outside stakeholders who might offer a valuable perspective, such as issue experts or impacted communities, are not always consulted. The result is that much of the AI currently being developed on a global scale is encoded with the goals, values, and assumptions of a narrow group of people.

Furthermore, many engineers, product managers, designers, and investors consider responsibility for AI to be outside the scope of their job. Growth- and profit-centered goals in the tech industry incentivize developers to collect as much data as possible and then figure out how to extract value from that data later. Unlike professions like medicine or civil engineering, software engineers are not required to take courses in ethics or get certified in standards for safety and reliability. Teaching students how to ask and explore ethical questions is one step forward — the next step is to empower tech workers to make changes in their workplaces.

6. Exploitation of workers & the environment

Vast amounts of computing power and human labor are used to build AI, and yet these systems remain largely invisible and are regularly exploited. The tech workers who perform the invisible maintenance of AI are particularly vulnerable to exploitation and overwork. The climate crisis is being accelerated by AI, which intensifies energy consumption and speeds up the extraction of natural resources.

Tech workers & labor

AI is developed and maintained by tech workers, who do not always have autonomy or power in their job. The development of AI has created a new class of tech workers who perform the invisible labor required to build and maintain these systems. While some AI systems are fully automated, most real-world tasks require some level of human discernment. "AI is simply not as smart as most people hope or fear,"[41] and much of what we call AI is a hybrid mix of human and machine collaborative decision-making. These workplace power differences are heightened for gig or contract workers, who are not considered employees of the company, but often rely on mobile apps and platforms to perform their work.

Companies building AI-powered services rely on a vast network of on-demand workers to clean and label datasets, and to train and improve models. Some of these on-demand workers use platforms like Mechanical Turk or Fiverr to perform different types of tasks. Many companies rely on their own set of contract workers to maintain their AI systems. For instance, when Amazon's Alexa trips up in a voice interaction, Amazon may send that information to a human worker who tags the interaction and helps improve the Alexa model. When Facebook flags possible hate speech or Twitter detects bot-like activity, information may be passed to a contractor to make a decision.

There are few employment laws globally that reflect the realities of the gig economy. This labor is often precarious and temporary, with few benefits or support. Workers who perform content moderation for platforms like Facebook and Twitter, for instance, are regularly subjected to disturbing imagery, sounds, and language, suffering serious mental health problems and secondhand trauma as a result.[42] When tech workers do decide to organize and speak out about their companies' business decisions, they run the risk of retaliation. At Google, Amazon, and Wayfair, tech workers have been fired or penalized for protesting their companies' contracts with US Immigration & Customs Enforcement (ICE). In order to build collective power among tech workers, we will need to continue exploring institutional and regulatory changes that empower tech workers within a precarious economy.

Environmental harms

While some AI implementations may help with understanding and monitoring the climate crisis, the current energy and resource demands of training AI models may well outweigh such benefits. Our natural resources are particularly vulnerable to exploitation and overuse, accelerating the already urgent global climate crisis. Over the past several decades, tech companies have driven higher levels of mining to produce computational devices.[43] In more recent years, AI development has spurred companies to collect increasingly large amounts of training data, resulting in unprecedented levels of energy consumption and expanding the need for data centers, which require space and enormous amounts of cooling resources.

AI optimizes the global extraction economy in ways we can't easily see or audit, speeding up extractive industries such as oil extraction, deforestation, and water management. Research suggests that AI is intensifying energy consumption, especially from the development of AI by major tech companies: Amazon, Microsoft, and Google.[44] Currently, there is little to no information about how much energy big tech's algorithms consume, but data suggest that the biggest carbon emissions are coming from training models and the storing of large datasets. The ad tech industry is assumed to be the biggest pollutant in this area.[45]

Tech companies continue to announce ambitious climate mitigation plans, often following pressure and mobilization from their workforce, but these efforts don't take full account of the harms caused by their AI systems – in terms of consumption, extraction, as well as social impact and community resiliency. What's more, AI and the climate crisis are displacing human rights, labor and land rights, and deepening racial inequalities. As such, any work to tackle AI's impact on the climate crisis must take an intersectional approach, bringing in voices and organizations from the racial justice, gender justice, environmental, and labor justice movements.

7. Safety and security

Malicious actors may be able to carry out increasingly sophisticated attacks by exploiting the vulnerabilities of intelligent systems.

Algorithmic curation is increasingly playing a role in information warfare as computational propaganda has become more sophisticated and subtle. AI can be used to surface targeted propaganda, misinformation, and other kinds of political manipulation. For instance, a *Washington Post* investigation[46] revealed that in the immediate aftermath of the 2018 Parkland school shooting in the US, people in online forums such as 8chan, 4chan, and Reddit developed a coordinated disinformation campaign to promote conspiracy theories. The campaign — which falsely portrayed the surviving Parkland students as "crisis actors" — was designed to mislead and divide the public over gun control.

Misinformation and disinformation are two terms used to describe this type of content. Misinformation refers to content that is characterized by its emotional impact and its potential to go viral. It is produced to propagate as widely and quickly as possible, even if the intention of

the creator wasn't malicious or to induce panic. Disinformation describes false content that is spread deliberately to mislead. Propaganda is not a new phenomenon, but what is new is the speed with which propaganda can be created and disseminated online, and manipulators' ability to target specific communities, groups, or individuals.[47]

Digital platforms create opportunities for a range of actors to exploit or "game" algorithmic systems for political or financial gain. Google's autocomplete suggestions, for instance, have been hijacked by malicious users to display antisemitic, sexist, and racist language.[48] Google Maps was once duped by a performance artist into displaying a traffic jam where there was none.[49]

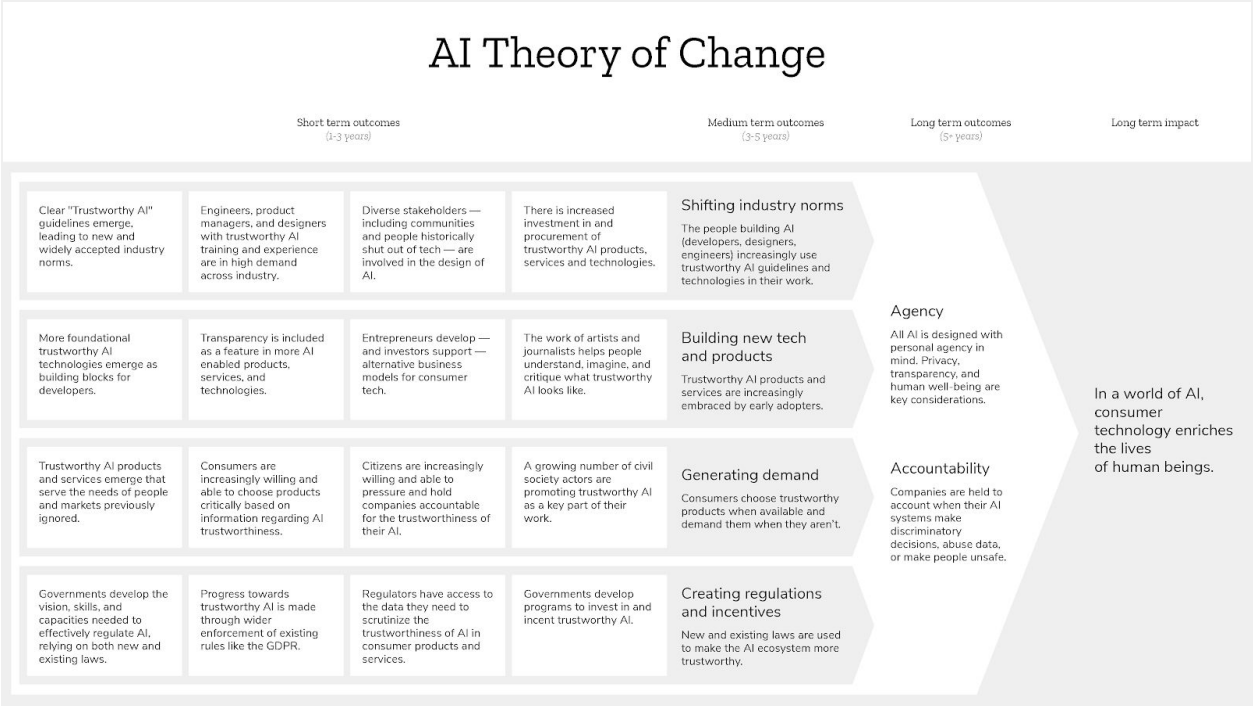
Manipulation of digital platforms is just one of several malicious uses of AI to which cybersecurity experts have said we may be particularly vulnerable.[50] AI can also be used to automate labor-intensive cyberattacks like spear phishing, carry out new types of attacks like voice impersonation, and exploit AI's vulnerabilities with adversarial machine learning.

In the physical world, AI can be used to hijack drones, self-driving cars, and other kinds of internet-connected devices. Hacks of Amazon's Ring doorbells that have been widely covered in the media, for instance, were powered by software that enabled hackers to automate brute force attacks on Ring accounts using a database of leaked usernames and passwords.[51] We have not yet seen an AI-powered cyberattack occur at scale, but cybersecurity experts are bracing themselves for the next wave of security threats.

V. The Path Forward

There is no way to chart a perfect course for addressing these challenges and developing more trustworthy AI — there are simply too many variables and forces at play. However, we can imagine the world we want to see, and draw a map for heading in that direction.

Sketching out this map was a major focus for Mozilla in 2019. As described in this paper's "README" section, we spent 12 months talking with experts, reading voraciously, and piloting AI-themed campaigns and projects. This exploration honed Mozilla's thinking on trustworthy AI by reinforcing several challenge areas including: monopolies and centralization, data governance and privacy, bias and discrimination, transparency and accountability, and industry norms. After looking at these challenges, we developed a theory of change that maps out how AI might look different and what steps we'd need to take to get there.



see Appendix A for full size diagram

This theory of change is not meant to describe the work that Mozilla alone needs to do — no one organization could possibly cover all this terrain on its own. Many other people and organizations will need to play a role in moving this agenda forward, from business leaders, investors, and academics to developers, policymakers, and people.

At the highest level, this theory of change posits that the technology that surrounds and shapes us should help us rather than harm us. Our long-term impact goal is:

> **In a world of AI, consumer technology enriches the lives of human beings.**

This might seem like an obvious statement, but it is not. The challenges we identified in the previous section illustrate how today's computing norms — and the underlying power differentials inherent to building today's versions of AI — both create opportunities and pose risks to billions of internet users. It's important to think critically about what it would look like to build technologies where benefits to humanity are at the forefront, and where we mitigate possible harms up front and by design. To do this, our theory of change focuses on two underlying principles:

1. **Agency: All AI is designed with personal agency in mind. Privacy, transparency, and human well-being are key considerations.**

2. **Accountability: Companies are held to account when their AI systems make discriminatory decisions, abuse data, or make people unsafe.**

These two principles are meant to work in tandem: one that is proactive, with a focus on creating more trustworthy tech from the design stage onward, and another that is defensive, recognizing that there will always be harms, risks, and bad actors that we need to defend against.

This reasoning is influenced by others working in the field as well as our own foundational thinking in the Mozilla Manifesto. The Manifesto's third entry — "The internet must enrich the lives of individual human beings" — was written in 2007, but hews close to Mozilla's current long-term impact goal. Mozilla's work has always focused on personal agency, privacy, and transparency (open source). The Mozilla Manifesto addendum, added in 2017, is also relevant to our current AI work: "We are committed to an internet that promotes civil discourse, human dignity, and individual expression."

In order to achieve more agency and accountability in AI in consumer-facing technologies, our theory of change outlines four medium-term outcomes that we should pursue:

1. **The people building AI increasingly use trustworthy AI guidelines and technologies in their work.**
2. **Trustworthy AI products and services are increasingly embraced by early adopters.**
3. **Consumers choose trustworthy products when available and demand them when they aren't.**
4. **New and existing laws are used to make the AI ecosystem more trustworthy.**

We'll need to see progress on all of these fronts to be successful. Success across multiple fronts was also necessary in the early 2000s, when control over the web was taken from Microsoft and put back in the hands of the public. Regulators reined in Microsoft's use of Windows to create an Internet Explorer monopoly. Open source developers created Firefox as a foundation for renewed web standards. Web developers latched onto these standards, making full-fledged cross-platform web apps like Gmail and Facebook the norm, and consumers flocked both to modern browsers like Firefox and these new web apps. The game had changed.

The good news is we are already seeing the seeds of changes like these in AI: developers wanting to build things differently, small companies developing new kinds of trustworthy products and technologies, people feeling suspicious about big tech, and regulators looking at ways to dismantle data monopolies.

Looking at these trends, this section of the paper provides high-level thinking on how we might collectively make progress on all four fronts. We also offer initial thoughts on the role that Mozilla might play in this broader work.

1. Shifting industry norms

Goal: The people building AI increasingly use trustworthy AI guidelines and technologies in their work.

Every era of computing tends to have norms that shape what is thought possible. We need to more clearly define what those norms are for AI, working closely with the people who are building AI, such as developers, designers, engineers, project managers, and data scientists.

At the core of this work will be efforts to ensure that people building AI are able to ask questions about responsibility and ethics at every stage in the AI research, product development, and deployment pipeline. We will need to support them on many levels — providing clear guidance, offering education and professional development, diversifying the workforce, and creating the right economic incentives. Importantly, all of this will need to get figured out iteratively and grow over time.

With this aim in mind, we believe that we should pursue the following short-term outcomes:

1.1. Best practices emerge in key areas of trustworthy AI, driving changes to industry norms.

Dozens of guidelines for "ethical AI" have been published in recent years, but they often focus on abstract, broad principles without clear steps for action. When they do point toward action, their ideas for operationalizing the principles vary widely across sectors. More work is needed to understand where these gaps appear.

A number of public and private sector organizations have published their own set of ethical principles to guide how AI systems are built and deployed. Prominent examples of these frameworks have been published by the EU's High-Level Expert Group, the Partnership on AI, the Organization for Economic Co-operation and Development (OECD), Google, SAP, the Association of Computing Machinery (ACM), Access Now, and Amnesty International. As this list suggests, these guidelines have come from all sectors of society — industry, government, and civil society.

Landscape scans of these frameworks published by Harvard's Berkman Klein Center[52] and Nature Machine Intelligence[53] show that across different sectors, there is global convergence around common principles such as transparency, fairness, and human well-being. According to one analysis of 84 frameworks, the most common principles included were transparency (86.9% of frameworks), justice and fairness (81.0%), a duty not to commit harm (71.4%), responsibility (71.4%), privacy (56.0%), and human well-being (48.8%).[54]

Most guidelines generally agree on core principles, but there are major differences across sectors about what they mean and how they should be implemented. For instance, in their assessment of transparency, nonprofits and data controllers tend to propose audits and oversight, whereas industry players propose technical solutions to transparency.[55] These fault lines show how the same sets of principles might result in different — and sometimes oppositional — actions when they are put into practice.

As of now, few players have prescribed concrete steps that translate their guidelines into action on the ground. This is because many of the challenges raised by AI are domain-specific and will require a coordination of work across sectors. How do developers interpret the guidelines in everyday technology and product development? What is the best way to ensure models are fair and unbiased? What steps must companies take to safeguard against harm? How do companies come to the difficult decision that an AI system should not be deployed at all? While we certainly need clear frameworks or guidelines to define what "good" looks like, we also need a concrete plan for putting those principles into action.

Our theory of change is meant to be a small step toward this action, a suggestion of what to do at a systems-wide level based on the challenges we face in consumer tech. Over the coming years, we want to work with people to put broad principles into action through AI development checklists, education programs, and software tooling. Putting these principles into action is key to shifting the norms of how AI is developed.

1.2. Engineers, product managers, and designers with trustworthy AI training and experience are in high demand across industry.

Engineers, product managers, designers, and other members of the cross-functional teams building AI wield a great degree of decision-making power. There are many initiatives underway aimed at helping developers think critically about their work, such as Mozilla's Responsible Computer Science Challenge[56], but more work needs to be done to ensure the people who are building AI responsibly are in high demand from companies.

One way to get there is to start with students before they have joined the workforce. A controversial New York Times op-ed[57] argues that academics have been "asleep at the wheel" when it comes to teaching ethics in tech. Indeed, the traditional approach to ethics education in computer science is far removed from engineers' day-to-day experience. Students say they don't connect with case studies and don't always know how it applies to their work. Further, many initiatives focus only on CS/Eng students when they should broaden to include other disciplines like information science, design, and management education.

Some universities have moved toward making ethical computing courses required for CS/Eng students, and also making these courses more practical. In a recent landscape analysis of 115 university courses in tech ethics, researchers conclude that while CS as a discipline has been slow to adopt ethical principles, it has made a great deal of progress in recent years. They recommend that students hear the message that "code is power" when they first start learning how to code and that this message should be reinforced throughout coursework.[58] More people are calling for major overhauls of how such degrees are taught altogether. At MIT, the New Engineering Education Transformation (NEET) group has been pioneering an alternative approach that teaches ethics as a set of skills embedded in what engineers already know.[59] Such classes train the next generation of AI developers to think not only about how they should design AI in future jobs, but also whether those systems should be built at all.

While promising, many of these initiatives are directed only at CS/Eng students and we have yet to see parallel efforts to integrate ethics into management or design education. Furthermore, these initiatives tend to focus on university coursework and don't include other forms of training, including online education or coding bootcamps. In order to fundamentally change how cross-functional teams build AI, we will need to think more broadly about all the different occupational roles that shape the development pipeline and ensure that they are empowered to think critically about their work. We will also need to make sure that people who follow non-traditional pathways into AI development are trained to ask critical questions about tech.

Companies are starting to recognize that in order to recruit and retain top talent, they will need to meet the rising demand for ethical tech, but we still have a long way to go. Skilled engineering graduates are already highly sought after and have more pull over potential employers than in many other industries. We aim to get to a point where tech companies are under pressure to demonstrate that they are building technology responsibly in order to attract top talent across disciplines – design, engineering, and management.

At the same time, there needs to be a major shift in company culture so that employees who are advocating for more responsible AI practices feel supported and empowered. Evidence suggests that the actions of internal advocates won't have impact unless their work is aligned with organizational practices.[60] We think that these two components — critically minded engineers and organizational change — need to be in place in order to usher in system-wide changes.

1.3. Diverse stakeholders — including communities and people historically shut out of tech — are involved in the design of AI.

It's not just the mindset of decision-makers that matters, but who is making those decisions. Tech has made strides in recent years to bring in new and diverse voices into product development, but we are still far from where we need to be.

The diversity crisis in the tech industry — and its direct link to problems with bias in AI — has been well documented. It has been reported that women make up only 10 percent of people working on "machine intelligence" at Google and 15 percent of Facebook's AI research group.[61] "It is not just that AI systems need to be fixed when they misrecognize faces or amplify stereotypes," says a recent AI Now report. "It is that they can perpetuate existing forms of structural inequality even when working as intended." It is crucial that the teams building AI are themselves diverse and represent a range of communities and perspectives.

Creating more diversity within developer communities, and specifically in who gets AI jobs and training, has a huge impact on how technology is built. A 2014 NCWIT report[62] found that gender-diverse management teams performed better in terms of overall productivity and team dynamics. Further, the study found that companies that dominated the market did so by encouraging innovation that drew from a diverse knowledge base. Diverse teams are more likely to drive innovation and change.

Engineering teams should strive to reflect the diversity of the people who use the technology, along racial, gender, geographic, socioeconomic, and accessibility lines. Those team members will be better attuned to the ways bias and discrimination manifest and they would also have a higher level of cultural context for how technologies might be received or interpreted in their community, region, or language.

Critically, changing the diversity of the people building AI will require making drastic changes to company culture. In its analysis of the diversity crisis in AI, AI Now concluded that a worker-driven movement aimed at addressing inequities holds the most promise for pushing for real change.[63] Companies must foster an open culture in which the status quo can be questioned or challenged without fears of retaliation.

Companies will need to develop processes for consulting with diverse communities throughout the AI product life cycle, especially when the technology may have an adverse impact on a historically marginalized community or region. This will require teams to adopt a more participatory, open approach to how it does its work, using frameworks and tools such as participatory design, co-design, or design justice.[64] It may also require companies to adopt stricter rules to safeguard against harm. Companies may mandate that particular features should be thoroughly tested with diverse user groups across geographic regions and languages before being deployed. Compliance frameworks exist for assessing risk and mitigating harm in AI, particularly when it comes to fairness and bias. But more work is required to ensure that companies not only adopt these narrow harm reduction processes, but proactively develop new tools and processes for designing and deploying AI.

1.4. There is increased investment in and procurement of trustworthy AI products, services and technologies.

Although there has been a rise in "impact investments" in socially responsible companies and startups, there is still a lot of work that needs to be done to ensure trustworthy AI products are getting the funding they need to become viable.

Impact investing — also known as socially responsible or ethical investing — is an investment approach that focuses on organizations and companies that are having a positive social impact on the world. It's a strategy that seeks to bring about environmental or societal change through investment. Impact investing represents a huge slice of investments in the US: The Forum for Sustainable and Responsible Investment estimates that in 2018, $12 trillion was invested in socially responsible investment funds, which represents 25% of all professionally managed assets in the US.[65] Evidence shows that young investors overwhelmingly want to invest in socially responsible companies.[66] Similarly, the B-corp movement in the UK legally requires B-corps to "consider the impact of their decisions on their workers, customers, suppliers, community, and the environment."[67]

In tech, this wave of impact investing is increasingly shaping what kinds of companies get funded. We are already seeing tech investors pay more attention to data privacy, a cornerstone of developing AI responsibly. Nearly $10 billion was invested in privacy and security companies in 2019, with the largest rounds of funding going to startups like Rubrik, 1Password, and OneTrust.[68]

We are also seeing larger tech companies pay more attention to privacy in their acquisition strategy, which is a step towards more trustworthy AI. In 2018, Apple acquired the privacy-conscious AI startup Silk Labs, which is building on-device machine learning software,[69] and Cisco acquired security startup Duo.[70] In 2019, Microsoft acquired BlueTalon, a data privacy and governance service.[71] Privacy is rapidly becoming a key part of a target company's risk profile in any acquisition.

There is a clear opportunity now for such "impact investors" who care about building tech responsibly to shape the AI product landscape. Most recently, 1,356 AI startups raised over $18.5 billion in funding in 2019, a new annual high for the AI sector.[72] VC funders and angel investors themselves attract huge amounts of capital if they focus their portfolios on AI. Impact investors should build on the momentum that has been growing in recent years around responsible tech and continue to fund AI startups that are building tech ethically.

Similarly, big tech companies looking to acquire AI have a huge amount of power. By acquiring socially responsible startups and technologies, these companies can send signals that building AI responsibly is not just a plus, but not doing so could be a major liability.

2. Building new tech and products

Goal: Trustworthy AI products and services are increasingly embraced by early adopters.

Introducing more trustworthy AI products on the market will require a number of things to happen. Foundational trustworthy AI technologies and practices — things like edge processing that use data locally on a device, machine learning techniques that require less data, or data trusts that balance power between companies and users — will need to emerge as building blocks for developers. Similarly, new thinking about business models, explainability, and transparency will be needed.

Importantly, we will also need to build up our imagination of how the products and services that are so central to our lives today can give users greater control over their digital lives. This aspect of the work will not only require the efforts of people developing products and services — startup founders, entrepreneurs, funders — but also journalists, computer science researchers, artists, and others who see the big picture of how things need to change.

With all of this in mind, we believe that we should pursue the following short term outcomes:

2.1. More foundational trustworthy AI technologies emerge as building blocks for developers.

A first major step towards better products and services is developing technological building blocks that can power more responsible AI. These building blocks — which could include better pre-trained models, alternative data governance models, privacy-preserving methods for machine learning, and decentralized, open source datasets — will reflect some of the trustworthy AI themes we've identified.

Tech infrastructure

One area where we are seeing better building blocks emerge is in the realm of privacy-preserving AI. Traditionally, machine learning needs centralized data access, which raises concerns about both privacy and centralization of control in the hands of a few large players. Increasingly, however, computer scientists have been exploring the possibilities of edge computing, decentralized computing that is done at or near the source of the data itself.

Researchers at Google have pioneered a method called **federated learning** that allows engineers to train a machine learning model without needing access to a centralized repository of training data.[73] Federated learning works by using a decentralized network of nodes or servers (i.e. people's devices) to train an AI system. Federated learning has the ability to

compute across millions of individual devices and combine those results to iteratively train the model. The open source library PySyft is a popular implementation of federated learning.[74] In this way, all the training data stays on a user's device or is split across a number of trusted servers rather than in a centralized database, ensuring greater privacy.

Another relevant technique, **differential privacy** is a statistical assessment of risk, in which data is statistically anonymized before being used to train the model so that individual users can't be identified from their data. Differential privacy has long been the foundation of Apple's approach to machine learning[75], and now there is an implementation of Google's open source tool TensorFlow for training machine learning models with differential privacy.[76]

Training datasets and pre-trained models are key to building AI. Although big tech companies might have the vast resources required to develop their own machine learning models, most smaller companies rely on datasets or pre-trained models from Google or IBM to develop their own AI applications. To this end, we will need to work toward ensuring that these existing pre-trained models and datasets are trustworthy. This will be complicated, as such a move risks further legitimizing the power of dominant tech players and further entrenching power inequalities. But by ensuring that the most popular pre-trained models and datasets adhere to a high level of scrutiny, we can quickly scale more powerful trustworthy AI in consumer tech.

Data governance

Another place where people are working on building blocks for more trustworthy AI is in the re-imagination of data governance and management. While regulations like the GDPR do this through the framework of individual data rights, we also need bottom-up legal structures that balance individual and collective approaches to data governance. Because "data ownership is both unlikely and inadequate as an answer to the problems at stake,"[77] we need to develop new approaches to data governance that shift power from companies back to communities and individuals through a mix of collective and individual frameworks. Ideally, individual fiduciary models can support and complement the more collectivized governance structures of data trusts and co-ops.

One proposed way to do this is to require the big tech platforms to become **information fiduciaries**.[78] A data fiduciary is an intermediary between individuals and data collectors, the people whose data is being collected ("data subjects") and the companies collecting that data ("data collectors"). Under this system, users would entrust their personal data to an online platform or company for a service, and in exchange the platform would have a responsibility to exercise care with that data in the interest of the user. While this approach holds promise, under this model companies would have a divided loyalty between the interests of shareholders and the interests of end users. According to Mozilla's report on alternative data governance models, "Critics are doubtful whether data fiduciary solutions present a realistic path to structural change, even if they could empower individuals to have more control over access to their personal data and enhance accountability through duties of care."[79]

An alternative fiduciary model is a legal mechanism called **data trusts**. Similar to an information fiduciary, a data trust is an independent intermediary between two parties. Unlike the fiduciary model, however, the trust would store the data from the data subjects and would negotiate data use with companies according to terms set by the trust. It would also have an undivided duty of loyalty toward its members according to a legal trust framework. Trusts could also serve as a mechanism for the collective enforcement of data rights, making it more likely that actions would be taken to use laws like GDPR to drive changes to products and services in ways that benefit end users. Different trusts might have different terms, and people would have the freedom to choose the trust that most aligns with their own expectations.[80] Some data trusts already exist. For instance, UK Biobank, a charitable company with trustees, is managing genetic data from half a million people.[81]

Another proposed approach is a **data cooperative** model. A data cooperative facilitates the collective pooling of data by individuals or organizations for the economic, social, or cultural benefit of the group. The entity that holds the data is often co-owned and democratically controlled by its members. Similar to a US credit union or a German *Sparkassen* savings bank[82], a data cooperative would share the benefits or profits of the data between its members. Because it could also run internal analytics, both data co-ops and data trusts would be in a strong position to negotiate better services for its members.[83] Some data co-ops already exist: The MIT Trust Data Consortium has demonstrated a pilot version of this system.[84]

Another approach to managing data differently is a **data commons**, which pools and shares data as a common resource and is typically accompanied by a high degree of community ownership and leadership. One of the major barriers to AI innovation is that only a handful of companies have access to training data. Data commons chip away at that power by democratizing access to training data sets and models, available to anyone who wants to use them and in a format that can be easily analyzed. Often that data is harmonized according to common data specifications, so that it is easy to use across different data pipelines. There are many different data commons already in existence: Mozilla's project Common Voice[85], for instance, is a crowdsourced dataset that represents the largest set of open source voice training data in the world, with more than 250k contributors, 4.2k hours of recorded voice data, and 40 different languages.[86] In academia, the UCI Machine Learning repository hosts hundred of datasets that have been accessed millions of times to benchmark ML algorithms in academic research.[87]

Privacy-preserving AI techniques, new data governance models, and open source training datasets are only a few examples of the kinds of building blocks we need to emerge in order to get closer to trustworthy AI. While some of the privacy-preserving techniques we mentioned are becoming standard, there is a long way to go before they are widely adopted. In addition, new data governance models and open data projects are still at a very early stage of development. Over the coming years, we plan to support people and organizations developing

and testing out these building blocks, starting with a major exploration of real-world uses of responsible data stewardship and data governance models.

2.2. Transparency is included as a feature in more AI enabled products, services, and technologies.

Transparency is the most commonly cited principle in dozens of ethical AI guidelines, across geographic regions and sectors[88] and a major focus of current research and development. There is wide consensus that technological norms and processes that enable transparency are themselves a major class of building blocks for trustworthy AI. However, different actors interpret transparency to mean different things. To move toward AI that is more transparent and accountable, we will need to weave together disparate work that is happening across different sectors.

The AI that is used in consumer-facing tech is often complex and opaque. There are a number of reasons for this, some of which are technical and some of which are based in institutional norms and incentives. According to Jenna Burrell, there are three key sources of AI opacity: (1) opacity as intentional corporate secrecy; (2) opacity as technical illiteracy; and (3) opacity that arises from the characteristics of machine learning algorithms and the scale required to apply them usefully.[89]

Focusing on that third category, technical solutions are currently being explored to make opaque AI more transparent. For decades computer scientists have been working to improve the explainability of AI — whether an AI system can be easily understood by and explained to a human. Another way developers are trying to make opaque AI more accountable is to use a human in the loop approach, which means that humans are directly involved in training, tuning, and verifying the data used in a machine learning algorithm.[90] This allows groups of experts with specialized knowledge to correct or fix errors in machine predictions as the process develops. In this way, humans are more actively involved in making normative judgements about the output of an AI system, rather than offloading decisions to the model.

However, technical solutions to AI transparency can only go so far. In order to address other sources of opacity such as intentional secrecy, companies will need to develop their products and services in a way that enables third party validation and audit.

While developers should be regularly auditing their AI systems,[91] they can also build those systems in a way that makes them easier to audit by third parties. One way companies are doing this is through open data archives. We are already seeing progress in this area with political ads: Companies that allow advertisers to target people with political ads should provide the public with clear, accurate, and meaningful information, available in a format that allows for bulk analysis by regulators and watchdogs groups. Under pressure from advocacy groups and policymakers, platforms like Facebook, Twitter, and Google have developed open political ad

libraries that provide detailed information about who paid for an ad, how it was targeted, the size of the audience that saw it, and other information.

The disclosure of data archives should go beyond advertising, though. In some cases, it may be necessary for companies to disclose what content was taken down or removed and why. As such, we want companies to work toward developing transparency products that give third parties access to more information about AI-based targeting systems. By analyzing data archives, researchers can more readily identify patterns of discrimination or deception that any individual user would not be able to see. Without bulk disclosure, the systems can evade efforts to systematically identify harm.

Platforms and services can be designed in a way that gives users greater control and agency over the algorithm's output. One way digital platforms are already doing this is by giving users explanations of the system's behavior within the UI itself. Such interventions can provide more information about inputs used by its recommender system. For instance, Netflix adds labels to its recommended videos with: "Because you watched X..." While such explanations may not help individuals identify discrimination or harm, they would help users understand why the algorithm behaved a certain way and could empower them to take action.

Transparency is not a "one size fits all" solution to AI accountability, and not every algorithm requires this level of transparency. In fact, arguments have been made that blunt transparency runs the risk of obscuring itself further by overwhelming us with too much information.[92] Furthermore, useful transparency isn't always possible – in which case we may need to reconsider whether AI should be used at all in high-stakes consumer environments like health or credit. But we do need to come up with real and workable solutions for transparency across the whole AI ecosystem — from transparency and control for users to tools that allow for scrutiny by researchers and regulators. As a community, we need to continue to experiment, test, and push to make this level of transparency a reality.

2.3. Entrepreneurs develop — and investors support — alternative business models for consumer tech.

We noted before that impact investors can set themselves apart by funding trustworthy AI products and technologies. In addition to new products, though, we will also need new business models that don't rely on the exploitation of people's data. As Zeynep Tufecki notes, a business model rooted in "vast data surveillance" to "opaquely target users...will inevitably be misused."[93] We need to ensure that startups with business models that are socially responsible and that don't exploit users are getting funded.

Companies already recognize the value in developing business models and objectives that respect — rather than infringe on — people's privacy. Over the past few years, data privacy has evolved from a "nice to have" for businesses to a must-have, critical topic of discussion.

According to a Cisco survey, 70% of organizations that invested in privacy say they are seeing benefits in every area of their business, in the form of competitive advantage, agility, and improved company attractiveness to investors.[94] Companies that demonstrate they care about people's privacy and well-being increasingly have a market advantage.

There is a hunger in the market for different business models that aren't focused on monetizing or exploiting people's data. One simple alternative is to set up the platform so that people pay to use it. Consumers are already used to this model — after all, many people are paying for access to streaming platforms like Netflix, HBO Go, and Hulu, or subscription services like Amazon Prime or HelloFresh. Before it was acquired by Facebook, WhatsApp was charging users $1 per year and the platform was still experiencing huge growth.[95] The downside to this model is that it's unclear whether people will be willing to pay for multiple subscription services. One survey shows that 75% of consumers capped their maximum spend on streaming services at $30.[96]

Platforms like Hulu and Facebook take what is called a two-sided approach[97] to their business model: They make profit both from (1) users paying to use the service, and (2) by sharing user data with advertisers. One option for two-sided businesses, then, is to rely on more privacy-preserving methods of doing data analysis. These new businesses may offer companies a new way to identify patterns without exploiting people's data.

It's important to note that even if these key industry business models were to change, companies may still be incentivized to exploit people's data for the purpose of training their models. This is the product of what has been called the "agile turn,"[98] a way of building tech that shortens development cycles and demands constant user surveillance and testing. More work needs to be done to change these incentives in the tech industry.

All of these ideas will have trade-offs. Funders will need to continue supporting startups and technologies that are seeking out different ways of doing business. As a way to build momentum, foundations and others interested in impact investing could set up special funds to encourage data privacy and alternative approaches to data governance. Impact funds have done a great deal to pave the way for investment in fields like green energy. The same could happen in the field of trustworthy AI.

2.4. The work of artists and journalists helps people understand, imagine, and critique what trustworthy AI looks like.

Many of AI's shortcomings are not readily apparent to the public, as they are often hidden in complex systems that are difficult to audit. The job of critiquing AI often falls on journalists, artists, creative technologists, and other researchers who are interrogating how these systems work. But beyond critique, they can help us expand our thinking around what is possible by showing us what alternate, preferred futures our technologies can offer.

Investigative journalism is shaping the future of AI by shedding light on technology's shortcomings and limitations. Journalists can serve as corporate watchdogs by investigating computational systems, and they can also help us understand what is happening by providing context and evidence. For instance, ProPublica's groundbreaking 2016 series on machine bias in crime algorithms[99] unlocked a new set of investigations into AI bias. In addition, its work on Facebook's ad targeting platform showed how advertisers could target ads in a discriminatory way, leading to new research and lawsuits on ad discrimination.[100] The fledgling news organization The Markup launched in 2020, using data to investigate tech and its influence on society.[101] The New York Times has a growing tech beat, hiring reporters with expertise in tech investigations. We rely on journalists to do the ever-important work of holding our technologies accountable when they have the potential for harm.

Similarly, art is helping expose the limitations and shortcomings of AI. Artists critique current systems and imagine different ones by providing us a new lens through which we can see our world. For instance, a project *ImageNet Roulette* by artist Trevor Paglen and Kate Crawford[102] exposes the biases in image datasets that are trained to categorize humans. The project reveals how AI can become a new vector for social discrimination, and shows how art can be wielded to hold tech accountable.[103]

In addition, art and design are tools to help us see what alternative worlds and technologies could look like. Artists and designers do this through speculative design and futures, tools that help us imagine the futures we want to build. For example, there is a growing body of work around feminist technologies that imagines what alternative voice AI could look like in practice. The organization Feminist Internet runs a workshop called "Designing a Feminist Alexa,"[104] which has resulted in a number of voice experiments that push the boundaries of how we think our voice assistants should speak, act, and interact.

As much of this work is still nascent, we have yet to see many of these experiments mirrored in our technologies just yet. Much more work will need to be done to ensure that these innovative ideas and experiments can become viable and real. Mozilla will continue to support journalists and artists who are critiquing our current technological landscape, and offering up visions of an alternate, preferred one.

3. Generating demand

Goal: Consumers choose trustworthy products when available and demand them when they aren't.

So far we have discussed shifting industry norms and developing products and services that integrate trustworthy AI. Ultimately, the impact and long term viability of efforts in these areas depends on consumer demand: Will the public support companies that protect their privacy?

Will people choose products or services that use AI responsibly? And are people ready to switch platforms, deactivate their accounts, or otherwise protest to demand better options? There is reason to (cautiously) believe the answer is "yes." A recent Cisco survey of consumers[105] reveals 84% of people care about data privacy and want more control over their data. More importantly, 32% are willing to act and have done so by switching companies or providers over data or data-sharing policies. This group tends to be younger, more affluent, shop more online, and are "early tech adopters" — a prime audience that you would think companies would be trying to cater to. Consumer surveys increasingly show that there is a market appetite for enhanced privacy and data protection.

Despite consumer interest, there are still significant barriers to generating the level of demand that would enable more trustworthy AI. Entrepreneurs developing products focused on privacy have a hard time reaching people, and people have little by way of reliable information to understand what products and services to trust, or whether they should trust a technology at all. People are confronted with complex terms and conditions, consent pages, and privacy settings that are difficult to sort through, even among the most tech-savvy. At the same time, large, established tech companies don't have a market incentive to build their AI differently, since they tend to lock people in. Consumer pressure could help change this.

With the aim of overcoming these barriers, we believe that we should pursue the following short-term outcomes:

3.1. Trustworthy AI products and services emerge that serve the needs of people and markets previously ignored.

A key step towards a broad market for trustworthy AI is the creation of products and services that meet the needs of people who are hungry for "something different." This includes people who want data privacy, a market that was considered marginal for many years. It also includes people whose interests, culture, communities, or life situation are not well served by existing AI and automated systems.

On the privacy front, we are starting to see a wave of startups whose core focus is bringing technologies like federated learning and differential privacy into consumer internet services. For instance, Owkin is an AI healthcare startup that uses a secure, federated framework in order to protect the sensitive data of patients.[106] Recently acquired by Sonos, Snips is an AI voice platform for connected devices that uses federated learning to do voice processing on devices, which protects user privacy. There is a great degree of skepticism, however, as to whether Sonos will invest meaningfully in privacy.

We've also started to see hints that established big tech players want to tap into the market for privacy — and that they are willing to integrate trustworthy AI technology as a part of this. Apple, for example, has made an extensive push to position itself as privacy friendly, with ads

saying, “What happens on your iPhone stays on your iPhone.”[107] This marketing pitch is backed in part by the use of both differential privacy and federated learning in core products like Siri, keeping voice samples on a user’s device unless they opt in to share them with Apple.[108] While the solid connection between marketing and privacy-preserving technology is laudable, it’s worth noting that the opt-in aspect of this privacy promise was only added after public outcry over Apple employees listening to people’s conversations as part of Siri’s training.[109] Even when the underlying technology lends itself to privacy, seemingly small decisions — like making “opt in’” the default setting — can make a big difference to people.

Of course, people seeking privacy are not the only people who have been ignored as AI has become central to internet products and services. People who speak non-dominant languages or who use non-Latin characters have historically been left out in products. For instance, the Amazon Alexa voice assistant speaks 15 different languages.[110] This may sound like a big number, but it’s tiny when compared to the 299 languages offered by Wikipedia.[111] Mozilla’s own Common Voice project[112] aims to be a counterpoint to the limited language offerings of voice tech like Alexa. It uses a Wikipedia-like, crowdsourcing approach to invite people to create voice training data in their own languages. It is the largest collection of open source voice training data in the world, with initial datasets in over 40 languages, including Catalan, Kabyle, Persian, Welsh, and Esperanto. It’s worth noting, though, that projects like Common Voice are still a long way from being integrated into consumer products and services that could meet the needs of a global audience.

While there are signals that companies will build products for people who want AI that is more trustworthy and inclusive, we still have a long way to go. Startups with this focus are few and far between and have a difficult time reaching their target markets. Open source initiatives aimed at inclusion and privacy have yet to make it into accessible mainstream products, and efforts by the big platform companies to serve markets like people seeking privacy are mixed at best. They still require significant investment — and rigorous scrutiny from governments, journalists, and people themselves.

3.2. Consumers are increasingly willing and able to choose products critically based on information regarding AI trustworthiness.

As more products using trustworthy AI reach the market, people will need better information about who and what to trust. At the moment, consumers don’t feel they can make educated choices about what products to buy or platforms to use. The Cisco survey mentioned previously[113] revealed that 43% of respondents believe that they aren’t able to protect their personal data. Of those who were worried, 73% said it was too hard to figure out what companies are actually doing with their data and 49% felt that they had no choice but to accept how their data was being used. People want greater transparency and agency, but they don’t have a way to get it.

There are a number of efforts to help consumers better understand the trade-offs between different products. Mozilla's *Privacy Not Included Guide[114] is a lightweight effort of this nature, providing people with plain-language reviews of AI voice assistants and other connected devices. The reviews include a rating against a set of minimum security standards as well as an analysis of how user data is treated. For instance, the review of the Facebook Portal voice assistant notes: "...data about your Portal usage — how often you do video calls, what apps you open, what features you use — can be used to target you with advertisements across Facebook. The company may also share specific demographic and audience engagement data with advertisers and analytics partners."[115] Information like this can be helpful to people choosing between different devices and services.

There are also efforts underway to develop more rigorous testing and labeling schemes, similar to nutrition labels on food products. Some early initiatives — such as the Harvard- and MIT-based Data Nutrition Project[116] — are aimed at helping data scientists and developers make their AI more trustworthy. Other projects are emerging to test whether products in the market are trustworthy and to help people make better choices. One example is Consumer Reports' Digital Standard[117] initiative, which looks at various criteria: encryption, potential overreach in the use of consumer data, and the transparency of the product's business model. While platforms like this have huge potential to empower people, they may be years away from being available to the public.

Reliable, easy-to-read information about "what's inside" AI-driven products and services will be essential if we want a more trustworthy AI ecosystem. It's important to recognize that efforts to provide this kind of information are still nascent. Not only is the amount of information available incredibly limited, but questions also remain as to what kind of information will be useful to people. Significant effort and funding will be needed in coming years to make the kind of progress that is necessary in this area.

3.3. Citizens are increasingly willing and able to pressure and hold companies accountable for the trustworthiness of their AI.

As we wait for clear consumer protection regulations or a mature market for trustworthy AI products and services to emerge, people will need to pressure companies directly to make the products they already use today more trustworthy.

There is a long history of this sort of consumer activism, where people want changes to how a product works or how it is made. An example of this type of activism is the Nike sweatshop campaigns of the 1990s.[118] Recognizing that Nike was contracting out to sweatshops and yet consumers would still buy Nike shoes, these campaigns focused on pushing the company to move to more ethical labor practices rather than boycotting its products outright. The campaign included both precise asks for changes in working conditions and for ongoing monitoring to ensure changes were maintained in factories on the ground. Campaigns of this nature have

become a regular part of consumer activism and are increasingly taken seriously by companies seeking to maintain a good reputation with the public.[119]

The ubiquity and near-monopoly status of companies like Facebook, Google, and Amazon make them good candidates for this kind of consumer pressure. Many people want to or have to use the products these companies offer, but they also want to trust that these companies are acting responsibly. There is evidence that there is already a strong consumer protest movement: A 2019 study from researchers at Mozilla and Northwestern found that a surprisingly large number of web users (30% of respondents) have intentionally changed their use of a product from the five major tech companies in protest of the company's actions.[120] Direct consumer campaigns with precise asks for product changes is one way to pressure companies to change their practices.

The #DeleteFacebook campaign that followed the Cambridge Analytica scandal was in some regards an example of this kind of campaign emerging in the consumer internet space. The goal of the campaign was to get people to either stop using Facebook or to use it in a different way. A 2018 Pew study[121] found that up to 74% of American Facebook users adjusted their privacy settings, took a break from the site, or deactivated their accounts after the Cambridge Analytica scandal. 24% deleted the Facebook app from their phones altogether. While the #DeleteFacebook campaign may have influenced these choices, such actions don't seem to have impacted Facebook's bottom line nor did they trigger substantive changes to the company's privacy practices.

A closer corollary to the Nike campaigns might be efforts to get YouTube to stop amplifying misinformation and other harmful content. While investigating the misinformation ecosystem in 2018, researchers discovered that YouTube's recommendation algorithm was heavily promoting misleading and sensational videos on topics like vaccinations, climate change, and white supremacy. The algorithm was designed to optimize for "user engagement" signals like watch time, which means that users were prompted to keep watching videos (with ads) for long periods of time.[122] Researchers, journalists, and nonprofits like Mozilla called on YouTube to make changes to address this problem, which they eventually began to do in early 2019.[123] As journalists questioned the efficacy of these changes,[124] Mozilla continued putting pressure on Google and ran public campaigns pushing YouTube for greater transparency that would allow researchers to audit its recommendation algorithm.

After an investigation of GoodRx revealed that the drug discount company was sending customer data to 20 third-party companies, Consumer Reports rolled out a public campaign to pressure the company to change its privacy practices. The campaign succeeded: GoodRx stopped sending data about customers' prescriptions to third parties like Facebook, and is now rolling out new privacy tools for consumers.[125]

The idea of using direct consumer pressure to push tech platforms for more trustworthy AI and data practices is promising. It offers a way to call for rapid and specific changes to the way

services are implemented — something could take years through lawmaking. However, this technique is nascent and has only had limited impact. It would seem that the key ingredient for a successful consumer-focused campaign in the US would be to link it with direct complaints to a federal regulator, as was the case in 2019 when consumer groups called on the FTC to investigate Facebook for knowingly deceiving children.[126]

In any case, such efforts demonstrate that meaningful consumer pressure campaigns can be a marathon that requires continuous and sustained effort. Much more work — and broader collaboration — is needed in this area to see how it can contribute to the development of trustworthy AI.

3.4. A growing number of civil society actors are promoting trustworthy AI as a key part of their work.

Over the last 25 years, a number of public interest organizations have emerged to promote digital rights and a healthy internet. Many of these organizations focus on data protection and AI in recent years. As experts on technology's impact on society, these organizations have the potential to play a significant role in advancing trustworthy AI. However, as a nascent field, it is unlikely that these organizations will be successful alone. They will need to form alliances with more established organizations from other fields if they are to drive the kinds of changes we need.

The field of digital rights and internet health includes organizations like the Electronic Frontier Foundation, Privacy International, European Digital Rights, Access Now and, of course, Mozilla. Most of these organizations have taken positions on AI. For example, Access Now issued a series of reports in 2018 arguing that we need to enhance data protections and create special safeguards for the use of AI by both governments and companies.[127] And Privacy International has taken the position that "there is a real risk that the use of new tools by states or corporations will have a negative impact on human rights."[128] While not specifically dedicated to advancing trustworthy AI, organizations like these bring established constituencies of technically minded activists and citizens. They offer a solid foundation for building public interest momentum around data protection and other AI-specific challenges.

A new crop of AI-focused public interest organizations has also emerged. This includes research organizations like AI Now Institute in the US and AlgorithmWatch in Germany. AI Now has played a central role in defining the public interest debate in the US on issues like bias and discrimination in AI and tech worker organizing. AlgorithmWatch conducts technical research into algorithms, including an investigation into the use of AI in Germany's credit scoring. These new organizations bring valuable expertise to the field, shaping the overall debate and advising governments on AI.

This increased focus on AI's impact on society is a step forward — it has already resulted in governments and companies taking these issues seriously. However, it is likely that more established nonprofits will be needed in this space if we want to generate the research, pressure, and political will needed to pressure governments and companies to act.

One promising development is the increased focus on privacy, data, and AI in traditional consumer rights groups. For example, Consumer Reports — a US organization with a long history protecting and informing consumers on everything from food safety to seat belts to finance protections — launched its Digital Lab in 2019 to build consumer power in the digital economy.[129] Consumers International, a collection of 250 consumer groups in 120 countries[130] has begun an effort to arm its members with research and campaign materials related to responsible AI. With large constituencies and deep connections into the consumer protection divisions of governments, these organizations have the potential to be powerful allies to dedicated digital rights and internet health organizations.

Another promising development is increased interest by civil and human rights organizations in the ways in which AI will impact the communities they serve. For example, the American Civil Liberties Union (ACLU) is asking the critical question of whether AI is making us less free[131] and Color of Change ran a campaign pushing Facebook for a civil rights audit.[132] We are also starting to see digital rights groups and more traditional organizations work together to find common interest around AI issues. In 2018, Access Now and Amnesty International led a coalition of public interest organizations to develop the Toronto Declaration, a call for equality and non-discrimination in the age of AI.[133] As organizations like these turn their attention to AI, there is a chance to both deepen thinking on the human and societal impacts of tech and to engage new constituencies on these issues.

The good news is that a strong civil society movement is emerging to rally around issues like privacy, data protection, and trustworthy AI. However, we still need to develop strong alliances between digital rights organizations and more traditional, established social justice organizations. AI is transforming the nature of discrimination and marginalization in society. While the digital rights space has technical expertise, it often lacks existing relationships with the communities that are most impacted, while those organizations that do have these relationships — such as human rights organizations, refugee organizations, or other groups working on racial justice, criminal justice, and poverty — lack the technical expertise.

From enshrining civil rights to getting seatbelts in cars to protecting rainforests, civil society organizations play a central role in pushing governments and companies to protect our common interests. Building alliances between digital rights groups and groups from other public interest sectors is likely the most effective way to meet this need.

4. Creating regulations and incentives

Goal: New and existing laws are used to make the AI ecosystem more trustworthy.

The development of new technologies is outpacing regulation in many countries. This means that new AI products and systems are being tested out on millions of people without effective government oversight or governance. At the same time, many lawmakers are eager to enact new regulations to limit the power of tech companies. But questions remain as to whether or not those laws are technically grounded and effectively address the problems at hand.

To improve the trustworthy AI landscape, we will need policymakers to adopt a clear, socially and technically grounded vision for regulating AI. We will also need lawmakers to ensure that baseline consumer and privacy protections serve as a cornerstone of any AI regulatory regime. Policymakers will need to enforce or update existing laws and enact new ones in order to meet the rising challenges.

One note: We have intentionally chosen to focus on Europe's regulatory landscape as a model for what positive change might look like. This is because some of the most interesting developments in data protections and governance are happening in Europe, and we consider the EU approach to regulating AI to be the most mature and promising. Its lawmaking could serve as a model for other countries that produce AI technologies. However, questions remain as to whether the EU model can and should be the model for other countries around the world, especially since AI expertise is unevenly distributed globally. Do countries that are not producing AI but are still using it need different sets of rules? Do developing economies have the resources to enforce rules? Given the global focus of Mozilla's work, we will need to address these questions as we continue to develop this body of work. We invite global partners to contribute by surfacing positive models and examples from their countries of what effective AI governance should look like.

With this framing in mind, we believe that we should pursue the following short-term outcomes:

4.1. Governments develop the vision, skills, and capacities needed to effectively regulate AI, relying on both new and existing laws.

Skills and capabilities

Most lawmakers do not have the skills and resources they need to craft effective policy on AI or big tech in general. Adding to the problem, civil society organizations don't always have the technical capacity to do informed research on AI, limiting their ability to advise governments on key issues. The result is that policy debates related to AI and data are typically dominated by experts and lobbyists from big technology companies.

Some policymakers are hiring technologists to inform and shape tech policy, but they have limited budgets and may prioritize hiring staffers with other expertise over those with technical expertise.[134] A report commissioned by Ford Foundation and other members of the NetGain partnership on the flow of tech talent into public sector jobs[135] finds that it is difficult to convince technologists to switch careers to join government offices because they cannot offer comparable salaries and benefits.

Without staff who have technical expertise and can navigate the nuances of tech policy, AI policymaking risks being tilted in favor of industry voices. Tech companies have a disproportionate amount of power over how policy is made in countries like the US, and many of them promote the narrative that all AI is "inscrutable." Different types of algorithms offer different levels of transparency, and not every algorithm is a "black box." However, tech companies often use "the cultural logic of the 'complicated inscrutable' technology…to justify the close involvement of the AI industry in policy-making and regulation."[136] Industry players involved in policymaking are the same group pushing for more invasive data collection.

Policymakers are strengthening their capacity by working with more technologists. Some governments are developing AI-specific centers of expertise, such as the UK's Office for Artificial Intelligence.[137] Other governments have created departments like the US Digital Service[138] that enlist technologists to support the development of civic technologies and tools. Technologists in digital service centers could help advise policymakers on key tech and AI issues. An emerging field of "public interest tech"[139] — supported by nonprofits like Mozilla, New America, and Ford Foundation — has also enabled technologists to influence tech policy decisions through fellowships like TechCongress, which places tech experts in a one-year fellowship with members of the US Congress or Congressional Committees.[140] These programs aim to bridge the knowledge gap by temporarily offering much-needed tech expertise to congressional offices.[141]

There's evidence that policymakers are listening to technologists from civil society. But nonprofits don't always have the technical capacity and they are often up against tech lobbyists and experts representing the interests of big tech companies. In its work on political advertising and the 2019 EU elections, the EU Commission put pressure on platforms like Facebook, Google, and Twitter to be more transparent about political advertising, aided by the technical expertise of not for profits like Mozilla.[142] When those companies failed to meet their commitments to the EU Code of Practice on Disinformation,[143] Mozilla helped the EU Commission understand why[144] and worked with researchers to make product recommendations.[145] As policymakers get up to speed on these issues, we need to ensure that nonprofits with technical expertise are part of the conversation.

Policymakers know that they need tech expertise and are making strides by working with technologists. But there is still more that needs to be done. Areas to invest in the coming years include expanding cross-disciplinary university programs that combine public policy and tech, and growing the number of research institutions with a focus on AI. In addition, governments

should continue to invest in the creation of technology centers of expertise that can be used across departments and ministries. Steps like these will help policymakers develop the skills and capacity they need to more effectively regulate AI.

Vision

Many governments are working toward building more effective regulatory regimes, starting with the first step of articulating that vision. While the growing momentum in this area is promising, many questions remain about whether emerging policy visions will both address the intertwined challenges like bias, privacy, and the centralization of control in AI and provide a practical approach that can be put into action.

At the highest level, governments and countries are working together to develop global governance frameworks for AI. In 2019, 42 countries took a critical step when they came together to endorse a global governance framework on AI, the OECD AI Principles.[146] Subsequently, the G20 adopted a set of global AI Principles, largely based on the OECD framework. The G20 principles affirm that companies building AI must be fair and accountable, their decision-making should be transparent, and they must respect values like equality, privacy, diversity, and international labor rights.[147] While frameworks like these are promising first steps, we are still far from a global consensus on what governance should look like in practice.

At the same time, countries are putting together their own governance frameworks to shape policymaking. In the EU, trailblazing privacy regulations like the GDPR have already transformed how companies work with user data to build AI, and member states will continue enforcing those laws. In a 2020 white paper,[148] the European Commission lays out its vision for governing AI, recommending that companies be required to provide documentation that datasets are statistically fair and unbiased, documentation describing how the AI was developed and trained, and enable greater government oversight. The EU is focused on a risk-based approach to regulation, which would target “high risk” areas that have the greatest potential for harm, such as healthcare, immigration, or employment.

Compared to other countries’ frameworks, the EU’s vision for trustworthy AI is the most mature. It suggests that there's no one-size-fits-all approach to regulating AI, and that extra safeguards are needed for deploying and using AI in 'high risk' situations. However, ‘risk’ in this approach is defined as ‘risk to an individual’, which excludes a whole category of AI applications that pose major collective risks. While the EU covers these risks in separate pieces of legislation, other countries should contemplate imposing transparency, documentation, and auditability requirements on 'high risk' AI-applications that have an impact on our broader democratic processes and institutions.

In total, over 60 countries have articulated their own visions for AI.[149] In the US, the White House announced the American AI Initiative, which focuses on driving technological innovation

and standards to protect a competitive edge on AI. However, regulations in the US have largely failed to keep pace with innovation, with American companies capitalizing on self-regulation. In 2019, the Algorithmic Accountability Act was introduced, which aims to boost federal oversight of data privacy and AI. In the UK, the Lords Select Committee put together a 2017 report[150] suggesting five overarching principles for an AI code that reinforce data rights, transparency, and social good. China has also published principles for governing "responsible AI"[151] in 2019, with a focus on human well-being, fairness, inclusivity, and safety. Australia laid out eight AI Ethics Principles[152] that are voluntary and aspirational, intended to complement any AI regulations. In 2020, Singapore launched an updated version of its Model AI Governance Framework,[153] which lays out the government's mature vision for using data and AI responsibly.

As governments develop the skills and capacity they need to catch up with current AI innovations, they will be able to prescribe more technically grounded and effective visions for governing AI. As they do this, it will be important to look holistically at AI, including its role in internet technologies that shape the lives of most of their citizens. It will also be important that their regulations incent companies that come up with trustworthy approaches to AI. Some governments are much further along in this process than others, however, and we still have yet to reach a global consensus on AI governance.

4.2. Progress toward trustworthy AI is made through wider enforcement of existing laws like the GDPR.

Policy tends to lag behind innovation, and the AI landscape is changing rapidly year to year. Given the scope of the risks and challenges posed by AI, designing a regulatory regime that can address all of these issues may feel daunting. The good news is that policymakers are not starting from scratch: Existing laws and regulations that protect data rights can be wielded in a meaningful way to address many of the challenges outlined in this paper.

The GDPR, which came into effect in 2018, is a prime example of an existing regulatory framework that can be used to address issues surrounding AI. For example, the GDPR has been used to pressure companies into taking data security seriously: Massive fines have been levied against British Airways and Marriott for their data breaches, although they have since been reduced to very low amounts on appeal.[154] The GDPR has been used to tackle the surveillance economy and rampant data collection that powers AI. In 2019, Google was fined €50 million for not disclosing to its users how data is collected across its various services for the purpose of serving them personalized ads. The penalty was the largest GDPR fine to date.[155] Growing enforcement of the GDPR in areas like these has a downstream effect of making the AI developed by these companies more privacy-friendly and trustworthy.

There are also sections of the GDPR that relate more directly to AI, but they have not yet been applied and tested. For instance, Article 22[156] of the GDPR, "Automated individual decision-making, including profiling," says that decisions made without any human intervention

cannot be used to make choices that could have a "significant impact" on an individual. This means that an algorithm can't be used to automatically decide, say, whether someone is eligible to qualify for a loan. In addition, the GDPR does not explicitly say that citizens have a "right to explanation," but according to Article 22 people do have a right to obtain "meaningful information about the logic involved"[157] in an automated decision that could have legal or significant impact. This means if someone's loan application is rejected by a bank's software, the bank may be required to provide general information about the input data used by the algorithm, or the parameters set in the algorithm.[158]

According to the European Data Protection Board's interpretation of the law[159], the GDPR covers the creation of and use of most algorithms. GDPR provisions that may apply to AI include: the requirement that processing be fair,[160] the principle of data minimization,[161] and data protection impact assessments.[162] Fair processing might require companies to "consider the likely impact of their use of AI on individuals and continuously reassess it." However, it might be impossible for a company to identify AI bias or perform impact assessments if that AI system is not sufficiently transparent.

Privacy regulations aren't the only laws that can be applied to the tech landscape in order to strengthen safe innovation in AI. Antitrust laws could be applied to help spur competition and innovation in AI. Currently the market for AI is less competitive and innovative because only a handful of tech companies dominate. Moreover, AI can accelerate the dominance of the few: Big tech companies have greater access to data, which allows them to develop better AI, which then allows them to collect even more data.

In the EU, authorities have not shied from imposing fines on big tech companies based on competition law. Google was fined €1.5 billion for antitrust violations in the online ad market in 2019. Authorities say Google was imposing unfair terms on companies that used its search bar on their websites in Europe.[163] Recently, a renewed interest in antitrust laws among legal scholars and regulators alike has presented an opportunity to strengthen competition policy.

Privacy protection laws like the GDPR are being adopted around the world, with Kenya and California passing similar laws in 2019. At the same time, the countries that pass such laws often do not have independent and sufficiently resourced regulators to enforce them effectively. There is an opportunity to use these trends to drive a trustworthy AI agenda, but only if both government and civic actors take a proactive role. Organizations like the Digital Freedom Fund, a European impact litigation organization, or the ACLU in the US, could play a role in bringing forward relevant cases under data protection laws. Alternatively, data co-ops could form to collectively represent millions of people under a single umbrella, providing both a way to enforce data rights en masse. If we can make it happen, then aggressive, creative, and technically grounded enforcement of existing laws could be a way to move towards trustworthy AI.

4.3. Regulators have access to the data and expertise they need to scrutinize the trustworthiness of AI in consumer products and services.

As we've seen, privacy laws like the GDPR address many of the concerns people have around how companies are collecting and processing data. But such privacy regulations do not specifically describe how companies should make their AI more transparent and accountable to third parties, nor is such oversight mandated by law (yet). In order to mitigate potential harm, we will need to explore what kind of transparency should be required by regulators in order to audit AI systems.

One way to increase understanding of an AI system is through blunt transparency — sharing the algorithm's source code. Complete transparency has a number of limitations, however, as it often ignores systems of power, runs the risk of obscuring itself further by overwhelming people with too much information, and can promote a false sense of knowledge.[164] Calls for transparency often fall short unless paired with clear explanation and documentation mandates, along with clear mechanisms ensuring that this information will be used to hold the system accountable by different stakeholders.

Some companies regularly audit their own AI systems to ensure accuracy and flag potential risks, but thus far self-regulation has largely failed to mitigate harm. Under pressure from regulators, companies are now starting to build AI systems in a way that makes them easier to audit by third parties, such as researchers or government agencies. According to the EU's 2020 AI White Paper,[165] transparency in this context could mean many things: from opening up the training data of an algorithm, to documentation of a system's robustness or accuracy, to more detailed record-keeping on the training methods and normative decisions made to build the AI system.

Depending on the context, companies may be compelled to release information about a model's training data. Such information may include how the data was obtained, a description of why a particular dataset was selected, proof that the data meets safety standards, is sufficiently broad and unbiased, and personal data is protected. Companies may also be compelled to release detailed documentation about how the AI was designed, programmed, and deployed. Such documentation could include records on the programming of the AI: what traits or values the model was optimizing for, or what the weights were for each parameter at the outset. Documentation may also include records on the training methodologies, processes, and techniques used to build, test, and validate the AI systems.[166] It is important that documentation include explanations for *why* a dataset or method was selected — normative explanations are critical pieces of information regulators need to understand the AI development workflow.[167]

In some contexts, companies or platforms may be compelled to develop data archives or public APIs that researchers, journalists, and other watchdogs can use to study patterns of discrimination or harm. Previously in this paper, we talked about how platforms like Facebook, Twitter, and Google have developed open political ad libraries that provide detailed information about the advertisements appearing on its platform, a first step towards empowering third parties to audit the platforms. However, when Mozilla assessed Facebook's Ad API ahead of the 2019 EU elections, researchers told us that the API did not allow them to download machine-readable data in bulk, nor was the data comprehensive and up-to-date.[168] Such companies should provide clear, accurate, and meaningful information to researchers and governments about its use of AI, and should be held accountable by policymakers and third party auditors.

Much more work needs to be done to determine what effective transparency and oversight looks like for AI, and what kind of data different stakeholders will need for effective audits. Transparency is not an end in itself, but it is a crucial prerequisite for meaningful accountability of AI systems. Developers will need to build AI in a way that makes it easier to audit, and people and governments will need to put pressure on companies to provide the data required for audit. We want to see enhanced levels of transparency across the board for companies building AI: transparency in terms of detailed documentation, information about the source code and training data, normative explanations of how the system was built, and the release of data archives and libraries that help researchers study AI systems and governments hold them accountable. This is an area in which greater standardization and rulemaking is needed.

4.4. Governments develop programs to invest in and incent trustworthy AI.

As governments hone their vision of how to regulate AI, many recognize the need for policies and programs that boost investment in research and startups in this area. They are also looking for ways to use procurement guidelines to ensure governments use trustworthy AI and encourage the growth of responsible businesses. Investment and procurement both offer governments a way to proactively build up industry segments that reflect the values in their AI vision, a move that is just as important as regulating AI.

One way governments are investing in the trustworthy AI ecosystem is by developing an industrial policy that matches their policy goals and vision for AI. In 2018, the European Commission announced that it would boost its investment in AI to €1.5 billion by 2020 in order to keep pace with Asia and the US[169] and the German government announced it had set aside €3 billion for AI R&D.[170] More recently, new proposals from the Commission suggest that Europe may increase its investment in AI to over €20 billion and is seeking to create a single European market for data.[171] In the US, where private investment in AI is already high, the White House issued an Executive Order encouraging AI investment but no clear plan. China's government, on the other hand, is investing heavily in AI: The government's VC fund is planning to invest more than $30 billion in AI within state-owned companies. One Chinese state is

investing $5 billion in AI tech and another major city, Tianjin, is investing $16 billion in its local AI industry.[172]

Another way governments support trustworthy AI is by developing a procurement strategy that matches their strategic vision for AI. So far, the software used by government agencies has not always demonstrated the level of transparency and accountability we might expect from any public use of technology. City governments and government agencies are not able to properly assess the AI-enabled systems they want to procure, which has led them to invest in or buy "AI snake oil."[173] Because there are no clear rules about public oversight of tech vendor contracts, government agencies may procure and use tech that could impact millions of people without ever needing to notify the public.

Some governments have taken steps to create guidelines for government agency procurement of AI-powered tech. In the UK, the government published a "Guide to using AI in the Public Sector"[174] based on its Data Ethics Framework[175] to enable public agencies to adopt AI systems in a way that benefits society. These procurement guidelines aim to empower government agencies to buy trustworthy AI by helping them evaluate suppliers and establish rules for transparency. Recommendations include developing a strategy for addressing the limitations of training data and focusing on accountability and transparency throughout procurement.

In the US, New York City established the Automated Decision Systems (ADS) Task Force in 2018 to set up a process for reviewing the use of algorithms by city agencies. AI Now Institute developed a practical framework for city procurement of AI technologies in the form of Algorithmic Impact Assessments[176] that recommends that cities inform the public of any proposed procurement, conduct internal agency self-assessments to make sure the agency has the capacity to assess fairness and disparate impact, and give researchers and auditors meaningful access to the AI system once it's deployed.

On a more global scale, the Cities Coalition for Digital Rights, a coalition of 39 cities in the EU and the US, are taking steps to ensure that cities use technology in an open and transparent way. In its declaration[177], the coalition affirms several broad principles including the transparency, accountability, and non-discrimination of algorithms. This means that the public "should have access to understandable and accurate information about the technological, algorithmic, and artificial intelligence systems that impact their lives," and they should be able to "question and change unfair, biased or discriminatory systems." In the future, this coalition may serve as a testing ground for enacting better procurement standards and rules.

As we've illustrated, governments are in the process of developing their own procurement guidelines for AI, but these guidelines have largely not been implemented yet. One way to operationalize these guidelines is for government agencies to adopt them directly into the terms and conditions of procurement contracts. For instance, such contracts might require any AI-powered software to meet a gold standard in terms of transparency, auditability, and

fairness. They may also include rules for public notice and review of the technology. In this way, government agencies and cities can use their buying power to support trustworthy AI products.

This is an area in which governments are least developed and could be a major opportunity for growth. Governments should seek to align their visions and framework for trustworthy AI with their industrial investment and tech procurement policies, thus creating incentives for better technologies and companies to emerge to meet rising demand.

VI. Conclusion

The work required to shift from centralized, privacy-invading AI to an era of trustworthy AI that respects people can seem daunting, but it is essential. Fortunately, we know that this kind of shift is feasible. Two decades ago, a broad coalition of people succeeded at shifting personal and business computing away from a platform tightly controlled by one company and towards a more open, decentralized internet.

Several points in this paper can be distilled down to a few big takeaways. We need to transition from discussion to action on trustworthy AI. We need to mobilize not just engineers and regulators, but also everyday people, investors, and entrepreneurs. We need to make it easy and desirable for people to switch to services that are truly trustworthy, ensuring that companies aren't just "trust washing." Finally, we need to focus on not just the individual harms of AI, but also the collective harms — how these systems intersect with society at large.

Obviously, Mozilla (or any single entity) can't do all this alone. Driving this kind of watershed change requires that we both work collaboratively with a large movement of others, and pick specific areas where we think we can make a difference. This is exactly what Mozilla has decided to do as part of its commitment to promote trustworthy AI.

One of the specific areas that Mozilla will develop is new approaches to data governance. This includes an initiative to network and fund people around the world who are prototyping collective data governance models like data trusts and data co-ops. It also includes our own efforts to build useful AI building blocks that can be used and improved by anyone, starting with our own open-source text-to-speech efforts such as DeepSpeech[178] and Common Voice data commons.[179] There is a great deal of technical, legal and regulatory work ahead of us in these areas. However, we believe that new models of data governance have the potential to be as transformative in the next quarter century as open source software was in the last. If these new models can work at scale, they have the potential to shift the power balance, putting users and small developers on a much more level playing field with the big tech companies.

Mobilizing people is another area where Mozilla believes that it can make a difference. This includes continued efforts to provide people with information they can use everyday to

understand and assess the products and services they are using, as we have done with our annual *Privacy Not Included Guide.[180] It also includes organizing people who want to push on companies to make specific changes to their products and services, building on campaigns we've run around Facebook, YouTube, Amazon, Venmo, Zoom, and others over recent years. Our hope is that this approach can at once complement the messages of more strident digital rights organizations and give tech companies real input that they can act on.

Ultimately, the most important spot Mozilla can pitch in is to demonstrate what trustworthy AI products and services look like in action. Mozilla has included AI and data sovereignty as themes in a new set of product innovation programs that it is developing through the course of 2020. The goal of this effort is to find and grow internet technologies that have the potential to improve the dynamics of life online — bringing the values of the Mozilla Manifesto to the kinds of digital products and services that will shape our lives over the next 20 years.

As noted earlier, the particular spots that Mozilla chooses to focus on can only be a small part of shifting from the current era of AI to one that is more trustworthy.

A significant portion of the investment that Mozilla will make in trustworthy AI will be about growing the movement of people working on these issues — something we've already been doing for a number of years. This includes identifying diverse stakeholders who share our vision, and then giving those people and projects the resources they need to grow. Through Mozilla's Fellowships and Awards work, we're already collaborating with data scientists in Nairobi, AI policy analysts in Brussels, online advertising watchdogs in London, privacy activists in São Paulo, and dozens of others. We will use these funding programs, as well as our annual Mozilla Festival, to help grow and connect the movement of people around the world working on topics related to trustworthy AI.

Importantly, this work will also include efforts to collaborate with organizations and movements not traditionally focused on issues like internet health. We've already moved in this direction, collaborating with organizations like Greenpeace and Friends of the Earth on our efforts to push Facebook towards better political ad transparency in the 2019 EU elections. This will continue with future efforts to collaborate with the consumer movement around the world and human rights organizations in the Global South in future campaigning efforts. At the same time, we will aim to build bridges between our trustworthy AI work and the mainstream tech sector.

As we've noted above, moving towards trustworthy AI will require a major shift in the norms that underpin our current computing environment and society. That is a huge change, but it is possible. It happened before with the shift from Windows to the web, and there are signs that it is already starting to happen again.

In recent years, online privacy has evolved from a niche issue to one routinely on the nightly news and newspapers' front pages. Now, as a result, many developers build encryption into their products as a matter of course, and value-oriented applications like Signal and various

VPNs are popular. Landmark data protection legislation has passed in Europe, Brazil, California, and elsewhere around the world, and people are increasingly saying that they want companies to treat them and their data with more care and respect. All of these trends bode well for the kind of shift that we believe needs to happen. With a focused, movement-based approach, we can make trustworthy AI a reality.

Endnotes

1. "AI Now Report 2018," AI Now Institute, December 2018, https://ainowinstitute.org/AI_Now_2018_Report.pdf.
2. "High-Level Expert Group on Artificial Intelligence," European Commission, June 14, 2018, https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence.
3. "Open Letter: Facebook, Do Your Part Against Disinformation," Mozilla, February 11, 2019, https://blog.mozilla.org/blog/2019/02/11/open-letter-facebook-do-your-part-against-disinformation/.
4. Ashley Boyd, "28 Reasons Why YouTube Must Change" Mozilla (Blog), October 15, 2019, https://foundation.mozilla.org/en/blog/28-reasons-why-youtube-must-change/.
5. "Introducing 'Stealing Ur Feelings,' an Interactive Documentary About Big Tech, AI, and You," Mozilla, September 23, 2019, https://blog.mozilla.org/blog/2019/09/23/introducing-stealing-ur-feelings-an-interactive-documentary-about-big-tech-ai-and-you/
6. "Spotlight: Let's ask more of AI," Internet Health Report 2019, Mozilla, April 2019, https://internethealthreport.org/2019/lets-ask-more-of-ai/
7. Mark Surman, "Why AI + Consumer Tech?," Mark Surman, April 23, 2019, https://marksurman.commons.ca/2019/04/23/why-ai-consumer-tech/.
8. DeepSpeech, Mozilla, https://github.com/mozilla/deepspeech.
9. George Roter, "Sharing Our Common Voices – Mozilla Releases the Largest to-Date Public Domain Transcribed Voice Dataset," The Mozilla Blog, February 28, 2019, https://blog.mozilla.org/blog/2019/02/28/sharing-our-common-voices-mozilla-releases-the-largest-to-date-public-domain-transcribed-voice-dataset.
10. "Can Your Holiday Gift Spy on You?," Mozilla, November 20, 2019, https://foundation.mozilla.org/en/blog/can-your-holiday-gift-spy-you/.
11. "The Trust Opportunity: Exploring Consumers Attitudes to the Internet of Things," Consumers International and the Internet Society, May 2019, https://www.internetsociety.org/wp-content/uploads/2019/05/CI_IS_Joint_Report-EN.pdf.
12. "Techlash? America's Growing Concern With Major Technology Companies," The Knight Foundation, March 2020, https://knightfoundation.org/wp-content/uploads/2020/03/Gallup-Knight-Report-Techlash-Americas-Growing-Concern-with-Major-Tech-Companies-Final.pdf.
13. Amnesty International and YouGov, "Surveillance Giants: How the Business Model of Google and Facebook Threatens Human Rights," 2019, https://www.amnesty.org/download/Documents/POL3014042019ENGLISH.PDF
14. Nellie Bowles, "'I Don't Really Want to Work for Facebook.' So Say Some Computer Science Students," *The New York Times*, November 15, 2018, https://www.nytimes.com/2018/11/15/technology/jobs-facebook-computer-science-students.html
15. Sidney Fussell, "How an Attempt at Correcting Bias in Tech Goes Wrong," *The Atlantic*, Oct 9, 2019, https://www.theatlantic.com/technology/archive/2019/10/google-allegedly-used-homeless-train-pixel-phone/599668/
16. "OECD Principles on AI," OECD, May 22, 2019, https://www.oecd.org/going-digital/ai/principles/
17. Michael Kratsios, "Artificial Intelligence: Next Steps", OECD Forum Network, May 22, 2019, https://www.oecd-forum.org/users/262053-michael-kratsios/posts/49175-artificial-intelligence-next-steps
18. "White Paper on Artificial Intelligence: A European approach to excellence and trust," European Commission, February 19, 2020, https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf
19. John Isaza and Hannah Katshir, "Brazil Passes Landmark Privacy Law: The General Law for the Protection of Privacy," Business Law Today, April 24, 2020, https://www.americanbar.org/groups/business_law/publications/blt/2020/05/brazil-privacy-law/.

20. Anirudh Burman and Suyash Rai, "What Is in India's Sweeping Personal Data Protection Bill?," Carnegie India, March 9, 2020, https://carnegieindia.org/2020/03/09/what-is-in-india-s-sweeping-personal-data-protection-bill-pub-80985.
21. Amy Webb, *The Big Nine: How The Tech Titans and Their Thinking Machines Could Warp Humanity*, PublicAffairs/ Hachette, March 5, 2019.
22. Daisuke Wakabayashi, "Prime Leverage: How Amazon Wields Power in the Technology World," *The New York Times*, December 15, 2019, https://www.nytimes.com/2019/12/15/technology/amazon-aws-cloud-competition.html
23. David McLaughlin and Aoife White, "Google's Fitbit Deal Tests Merger Cops Eyeing Data Giants," *Bloomberg Businessweek*, February 10, 2020, https://www.bloomberg.com/news/articles/2020-02-10/google-fitbit-deal-poses-test-for-merger-cops-eyeing-data-giants.
24. "House lawmakers ask Apple, Amazon, Facebook and Google to turn over trove of records in antitrust probe," *The Washington Post*, September 13, 2019, https://www.washingtonpost.com/technology/2019/09/13/house-lawmakers-ask-apple-amazon-facebook-google-turn-over-trove-records-antitrust-probe/
25. Nick Srnicek, "We need to nationalise Google, Facebook and Amazon. Here's why," *The Guardian*, August 30, 2017, https://www.theguardian.com/commentisfree/2017/aug/30/nationalise-google-facebook-amazon-data-monopoly-platform-public-interest
26. Jack Nicas, "Atlanta Asks Google Whether It Targeted Black Homeless People," *The New York Times*, October 4, 2019, https://www.nytimes.com/2019/10/04/technology/google-facial-recognition-atlanta-homeless.html
27. Austin Carr and others, "Silicon Valley Is Listening to Your Most Intimate Moments," *Bloomberg Businessweek*, December 11, 2019, https://www.bloomberg.com/news/features/2019-12-11/silicon-valley-got-millions-to-let-siri-and-alexa-listen-in
28. Ginger Zhe Jin, "Artificial Intelligence and Consumer Privacy," SSRN Scholarly Paper (Rochester, NY: Social Science Research Network, January 1, 2018), https://papers.ssrn.com/abstract=3112040.
29. Joy Buolamwini and Timnit Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification," in *Conference on Fairness, Accountability and Transparency* (Conference on Fairness, Accountability and Transparency, PMLR, 2018), 77–91, http://proceedings.mlr.press/v81/buolamwini18a.html.
30. Safiya Umoja Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism* (New York, NY: New York University Press, 2018).
31. Tomiwa Ilori, "Facebook's Content Moderation Errors Are Costing Africa Too Much," *Slate Magazine*, October 27, 2020, https://slate.com/technology/2020/10/facebook-instagram-endsars-protests-nigeria.html.
32. Till Speicher et al., "Potential for Discrimination in Online Targeted Advertising," in *FAT 2018 - Conference on Fairness, Accountability, and Transparency*, vol. 81 (New-York, United States, 2018), 1–15, https://hal.archives-ouvertes.fr/hal-01955343.
33. Fairness, Accountability, and Transparency in Machine Learning, accessed May 11, 2020, https://www.fatml.org/
34. Zeynep Tufekci, "YouTube, the Great Radicalizer," *The New York Times*, March 10, 2018, https://www.nytimes.com/2018/03/10/opinion/sunday/youtube-politics-radical.html.
35. Finale Doshi-Velez and Been Kim, "Towards A Rigorous Science of Interpretable Machine Learning," ArXiv:1702.08608 [Cs, Stat], March 2, 2017, http://arxiv.org/abs/1702.08608.
36. M. C. Elish and Tim Hwang, "An AI Pattern Language," Data & Society, 2017, https://www.datasociety.net/pubs/ia/AI_Pattern_Language.pdf.
37. "Inspecting Algorithms in Social Media Platforms," Ada Lovelace Institute, November 2020, https://www.adalovelaceinstitute.org/wp-content/uploads/2020/11/Inspecting-algorithms-in-social-media-platforms.pdf.
38. Frederic Lardinois, "YouTube Changes Its Search Ranking Algorithm To Focus On Engagement, Not Just Clicks," *TechCrunch* (blog), October 12, 2012, https://techcrunch.com/2012/10/12/youtube-changes-its-search-ranking-algorithm-to-focus-on-engagement-not-just-clicks/

39. Christoffer Hernæs, “Is Technology Contributing to Increased Inequality?,” *TechCrunch* (blog), http://social.techcrunch.com/2017/03/29/is-technology-contributing-to-increased-inequality/.
40. Tom Simonite, “AI is the future - but where are the women?” *Wired*, August 17, 2018, https://www.wired.com/story/artificial-intelligence-researchers-gender-imbalance/.
41. Mary L. Gray and Siddharth Suri, *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass*, Houghton Mifflin Harcourt, 2019.
42. Casey Newton, “The Trauma Floor: The secret lives of Facebook moderators in America,” *The Verge*, February 25, 2019, https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona.
43. Fleur Scheele, Esther de Haan, and Vincent Kiezebrink, “Cobalt blues: Environmental pollution and human rights violations in Katanga’s copper and cobalt mines,” *SOMO*, April 2016, https://www.somo.nl/wp-content/uploads/2016/04/Cobalt-blues.pdf.
44. “Oil in the Cloud: How Tech Companies are Helping Big Oil Profit from Climate Destruction,” Greenpeace USA, May 19, 2020, https://www.greenpeace.org/usa/reports/oil-in-the-cloud/.
45. Roel Dobbe and Meredith Whittaker, “AI and Climate Change: How They’re Connected, and What We Can Do about It,” AI Now Institute, October 17, 2019, https://medium.com/@AINowInstitute/ai-and-climate-change-how-theyre-connected-and-what-we-can-do-about-it-6aa8d0f5b32c.
46. Craig Timberg and Drew Harwell, “We studied thousands of anonymous posts about the Parkland attack — and found a conspiracy in the making,” *The Washington Post*, February 27, 2018, https://www.washingtonpost.com/business/economy/we-studied-thousands-of-anonymous-posts-about-the-parkland-attack---and-found-a-conspiracy-in-the-making/2018/02/27/04a856be-1b20-11e8-b2d9-08e748f892c0_story.html.
47. Renee DiResta, “Computational Propaganda: If You Make It Trend, You Make It True,” *The Yale Review* 106, no. 4 (2018): 12–29, https://doi.org/10.1111/yrev.13402.
48. Samuel Gibbs, “Google Alters Search Autocomplete to Remove ‘are Jews Evil’ Suggestion,” *The Guardian*, December 5, 2016, sec. Technology, https://www.theguardian.com/technology/2016/dec/05/google-alters-search-autocomplete-remove-are-jews-evil-suggestion.
49. Brian Barrett, “An Artist Used 99 Phones to Fake a Google Maps Traffic Jam,” *Wired*, accessed September 16, 2020, https://www.wired.com/story/99-phones-fake-google-maps-traffic-jam/.
50. Miles Brundage et al., “The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation,” *ArXiv:1802.07228 [Cs]*, February 20, 2018, http://arxiv.org/abs/1802.07228.
51. Joseph Cox and Samantha Cole, “How Hackers Are Breaking Into Ring Cameras,” *Vice Motherboard*, December 11 2019, https://www.vice.com/en_us/article/3a88k5/how-hackers-are-breaking-into-ring-cameras
52. Jessica Fjeld, Nele Achten, Hannah Hilligoss, Adam Nagy, and Madhulika Srikumar. “Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI,” January 15, 2020, *Berkman Klein Center Research Publication* No. 2020-1, https://ssrn.com/abstract=3518482.
53. Anna Jobin, Marcello Ienca and Effy Vayena, “The global landscape of AI ethics guidelines,” *Nature Machine Intelligence*, vol. 1, no. 9, Sept. 2019, pp. 389–99, https://www.nature.com/articles/s42256-019-0088-2.
54. Ibid.
55. Ibid.
56. “Responsible Computer Science Challenge,” Mozilla Foundation, accessed April 28, 2020, https://foundation.mozilla.org/en/initiatives/responsible-cs/.
57. Cathy O’Neil, “The Ivory Tower Can’t Keep Ignoring Tech,” *The New York Times*, November 14, 2017, https://www.nytimes.com/2017/11/14/opinion/academia-tech-algorithms.html.
58. Casey Fiesler, Natalie Garrett, and Nathan Beard, “What Do We Teach When We Teach Tech Ethics?: A Syllabi Analysis,” in *Proceedings of the 51st ACM Technical Symposium on Computer Science Education* (SIGCSE ’20: The 51st ACM Technical Symposium on Computer Science Education, Portland OR USA: ACM, 2020), 289–95, https://doi.org/10.1145/3328778.3366825.
59. “3 Questions: Marion Boulicault and Milo Phillips-Brown on ethics in a technical curriculum,” *MIT News*, March 11, 2020, https://news.mit.edu/2020/integrating-ethics-technical-curriculum-0311.

60. Michael A. Madaio et al., "Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20 (Honolulu, HI, USA: Association for Computing Machinery, 2020), 1–14, https://doi.org/10.1145/3313831.3376445.
61. Stephanie Wykstra, "Developing a More Diverse AI," *Stanford Social Innovation Review*, January 2019, https://ssir.org/articles/entry/developing_a_more_diverse_ai.
62. "What is the Impact of Gender Diversity on Technology Business Performance?," National Center for Women & Information Technology, 2014, https://www.ncwit.org/sites/default/files/resources/impactgenderdiversitytechbusinessperformance_print.pdf.
63. Sarah Myers West, Meredith Whittaker, and Kate Crawford, "Discriminating Systems: Gender, Race, and Power in AI," AI Now Institute, accessed May 11, 2020, https://ainowinstitute.org/discriminatingsystems.pdf.
64. Sasha Costanza-Chock, *Design Justice: Community-Led Practices to Build the Worlds We Need*, Information Policy (MIT Press, 2020).
65. "US Sustainable, Responsible and Impact Investing Trends," US SIF Foundation, The Forum for Sustainable and Responsible Investment, October 31, 2018, https://www.ussif.org/trends.
66. Adam Shell, "Millennial 401(k)s: A Peek inside Their 'Socially Responsible' Investments," *USA TODAY*, https://www.usatoday.com/story/money/2018/05/11/millennials-socially-responsible-investing/580434002/.
67. "Certified B Corporation," accessed April 28, 2020, https://bcorporation.uk/.
68. Gené Teare, "Almost $10B Invested In Privacy And Security Companies In 2019," CrunchBase News, January 29, 2020, https://news.crunchbase.com/news/almost-10b-invested-in-privacy-and-security-companies-in-2019/.
69. James Vincent, "Apple reportedly buys AI startup with privacy-conscious approach," *The Verge*, Nov 21, 2018, https://www.theverge.com/2018/11/21/18106192/apple-privacy-ai-silk-labs-acquisition.
70. Ron Miller, "Cisco's $2.35 billion Duo acquisition front and center at earnings call," *TechCrunch*, August 16, 2018, https://techcrunch.com/2018/08/16/ciscos-2-35-billion-duo-acquisition-front-and-center-at-earnings-call/.
71. Frederic Lardinois, "Microsoft acquires data privacy and governance service BlueTalon," *TechCrunch*, July 29, 2019, https://techcrunch.com/2019/07/29/microsoft-acquires-data-privacy-and-governance-service-bluetalon/.
72. Chris O'Brien, "AI startups raised $18.5 billion in 2019, setting new funding record," *VentureBeat*, January 14, 2020, https://venturebeat.com/2020/01/14/ai-startups-raised-18-5-billion-in-2019-setting-new-funding-record/.
73. Brendan McMahan and Daniel Ramage, "Federated Learning: Collaborative Machine Learning without Centralized Training Data," Google AI Blog, April 6, 2017, https://ai.googleblog.com/2017/04/federated-learning-collaborative.html
74. Saransh Mittal, "Federated Learning with PySyft," Medium, October 9, 2019, https://towardsdatascience.com/federated-learning-3097547f8ca3.
75. "Apple Differential Privacy Technical Overview," Apple, accessed April 6, 2020, https://www.apple.com/privacy/docs/Differential_Privacy_Overview.pdf
76. Carey Radebaugh and Ulfar Erlingsson, Introducing TensorFlow Privacy: Learning with Differential Privacy for Training Data, TensorFlow (Blog), March 6, 2019, https://blog.tensorflow.org/2019/03/introducing-tensorflow-privacy-learning.html
77. Sylvie Delacroix and Neil Lawrence, "Bottom-Up Data Trusts: Disturbing the 'One Size Fits All' Approach to Data Governance," *SSRN Scholarly Paper* (Rochester, NY: Social Science Research Network, October 12, 2018), https://doi.org/10.2139/ssrn.3265315.
78. Jack M. Balkin and Jonathan Zittrain, "A Grand Bargain to Make Tech Companies Trustworthy," *The Atlantic*, October 3, 2016, https://www.theatlantic.com/technology/archive/2016/10/information-fiduciary/502346/.

79. Stefan Baack and Madeleine Maxwell, “Alternative Data Governance Approaches: Global Landscape Scan and Analysis,” Mozilla Foundation, September 2020, https://foundation.mozilla.org/en/initiatives/data-futures/data-for-empowerment/.
80. Jack M. Balkin and Jonathan Zittrain, “A Grand Bargain to Make Tech Companies Trustworthy,” The Atlantic, October 3, 2016, https://www.theatlantic.com/technology/archive/2016/10/information-fiduciary/502346/.
81. Jack Hardinges, “What Is a Data Trust?,” Open Data Institute, July 10, 2018, https://theodi.org/article/what-is-a-data-trust/.
82. Claudia Irigoyen, “Sparkassen Savings Banks in Germany,” Centre for Public Impact (CPI) (blog), March 27, 2017, https://www.centreforpublicimpact.org/case-study/sparkassen-savings-banks-germany/.
83. Thomas Hardjono and Alex Pentland, “Data Cooperatives: Towards a Foundation for Decentralized Personal Data Management,” *ArXiv:1905.08819 [Cs]*, May 21, 2019, http://arxiv.org/abs/1905.08819.
84. Alex Pentland, et al., “Data Cooperatives: Digital Empowerment of Citizens and Workers,” MIT Connection Science, January 2, 2019, http://ide.mit.edu/sites/default/files/publications/Data-Cooperatives-final.pdf.
85. Common Voice, Mozilla, accessed April 25, 2020, https://voice.mozilla.org/en.
86. “4200h Voice Dataset Release: More Than 4,200 Common Voice Hours Now Ready For Download,” Mozilla Discourse, January 14, 2020, https://discourse.mozilla.org/t/4200h-voice-dataset-release-more-than-4-200-common-voice-hours-now-ready-for-download/52013.
87. Dheeru Dua and Casey Graff, UCI Machine Learning Repository, University of California School of Information and Computer Science, accessed July 20, 2020, http://archive.ics.uci.edu/ml.
88. Anna Jobin, Marcello Ienca and Effy Vayena, “The global landscape of AI ethics guidelines,” *Nature Machine Intelligence*, vol. 1, no. 9, Sept. 2019, pp. 389–99, https://www.nature.com/articles/s42256-019-0088-2
89. Jenna Burrell, “How the Machine ‘Thinks’: Understanding Opacity in Machine Learning Algorithms,” *Big Data & Society* 3, no. 1 (January 5, 2016): 205395171562251, https://doi.org/10.1177/2053951715622512.
90. Mothi Venkatesh, “What Is Human-in-the-Loop for Machine Learning?,” *Hacker Noon*, July 23, 2018, https://hackernoon.com/what-is-human-in-the-loop-for-machine-learning-2c2152b6dfbb.
91. Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort, “Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms,” paper presented to “Data and Discrimination: Converting Critical Concerns into Productive Inquiry,” a preconference at the 64th Annual Meeting of the International Communication Association, May 22, 2014, Seattle, WA, USA, http://www-personal.umich.edu/~csandvig/research/Auditing%20Algorithms%20--%20Sandvig%20--%20ICA%202014%20Data%20and%20Discrimination%20Preconference.pdf.
92. Mike Ananny and Kate Crawford, “Seeing without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability,” *New Media & Society*, December 13, 2016, https://doi.org/10.1177/1461444816676645.
93. Zeynep Tufekci, “Facebook’s Surveillance Machine,” *The New York Times*, March 19, 2018, sec. Opinion, https://www.nytimes.com/2018/03/19/opinion/facebook-cambridge-analytica.html.
94. “From Privacy to Profit: Achieving Positive Returns on Privacy Investments,” Cisco Data Privacy Benchmark Study 2020, Cisco, January 2020, https://www.cisco.com/c/dam/en/us/products/collateral/security/2020-data-privacy-cybersecurity-series-jan-2020.pdf.
95. Zeynep Tufecki, “Yes, Big Platforms Could Change Their Business Models,” *Wired*, Dec 17, 2018, https://www.wired.com/story/big-platforms-could-change-business-models/.
96. “New Survey Points to Streaming Subscription Fatigue Among U.S. Consumers,” *theTradeDesk*, January 6, 2020, https://www.thetradedesk.com/press-releases/new-survey-points-to-streaming-subscription-fatigue-among-u-s-consumers.
97. Andrei Hagiu and Julian Wright, “Multi-Sided Platforms,” *SSRN Scholarly Paper* (Rochester, NY: Social Science Research Network, March 19, 2015), https://doi.org/10.2139/ssrn.2794582.
98. Seda Gurses and Joris van Hoboken, “Privacy after the Agile Turn,” *SocArXiv*, May 2, 2017, https://doi.org/10.31235/osf.io/9gy73.

99. Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, "Machine Bias," *ProPublica*, May 23, 2016, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
100. Ava Kofman and Ariana Tobin, "Facebook Ads Can Still Discriminate Against Women and Older Workers, Despite a Civil Rights Settlement," *ProPublica*, Dec. 13, 2019, https://www.propublica.org/article/facebook-ads-can-still-discriminate-against-women-and-older-workers-despite-a-civil-rights-settlement
101. The Markup, accessed May 11, 2020, https://themarkup.org/
102. Kate Crawford and Trevor Paglen, Excavating AI: The Politics of Images in Machine Learning Training Sets, accessed April 20, 2020, https://www.excavating.ai/.
103. Naomi Rea, "How ImageNet Roulette, an Art Project That Went Viral by Exposing Facial Recognition's Biases, Is Changing People's Minds About AI," *artnet news*, September 23, 2019, https://news.artnet.com/art-world/imagenet-roulette-trevor-paglen-kate-crawford-1658305
104. Designing a Feminist Alexa: An experiment in feminist conversation design," Feminist Internet, accessed May 11, 2020, http://www.anthonymasure.com/content/04-conferences/slides/img/2019-04-hypervoix-paris/feminist-alexa.pdf.
105. "Consumer Privacy Survey: The growing imperative of getting data privacy right," Cisco Cybersecurity Series 2019, Cisco, November 2019, https://www.cisco.com/c/dam/en/us/products/collateral/security/cybersecurity-series-2019-cps.pdf.
106. "Federated Learning," Owkin, accessed May 11, 2020, https://owkin.com/federated-learning/.
107. Vlad Savov, "Apple Trolls CES with a Giant Dig at Android and Alexa Privacy," *The Verge*, January 5, 2019, https://www.theverge.com/2019/1/5/18169781/apple-google-privacy-troll-billboard.
108. Karen Hao, "How Apple Personalizes Siri without Hoovering up Your Data," *MIT Technology Review*, accessed April 5, 2020, https://www.technologyreview.com/s/614900/apple-ai-personalizes-siri-federated-learning/.
109. Dieter Bohn, "Apple Was a Little behind on Siri Privacy, Now It's Way Ahead," *The Verge*, August 29, 2019, https://www.theverge.com/2019/8/29/20837077/apple-siri-privacy-opt-out-voice-human-grading-review.
110. "Amazon's Alexa Now Speaks Hindi," *TechCrunch* (blog), September 18, 2019, https://social.techcrunch.com/2019/09/18/amazon-alexa-hindi-india/.
111. "List of Wikipedias," *Wikipedia*, accessed on April 5, 2020, https://en.wikipedia.org/wiki/List_of_Wikipedias/.
112. "Common Voice," Mozilla, accessed May 11, 2020, https://voice.mozilla.org/.
113. "Consumer Privacy Survey: The growing imperative of getting data privacy right," Cisco Cybersecurity Series 2019, Cisco, November 2019, https://www.cisco.com/c/dam/en/us/products/collateral/security/cybersecurity-series-2019-cps.pdf.
114. "*Privacy Not Included: A Buyer's Guide for Connected Products," Mozilla Foundation, https://foundation.mozilla.org/en/privacynotincluded/categories/smart-home/.
115. "Facebook Portal - *Privacy Not Included: A Buyer's Guide for Connected Products," accessed April 5, 2020, https://foundation.mozilla.org/en/privacynotincluded/products/facebook-portal/.
116. The Data Nutrition Project, accessed May 11, 2020, https://datanutrition.org/.
117. The Digital Standard, accessed May 11, 2020, https://www.thedigitalstandard.org/.
118. B. J. Bullert, "Progressive Public Relations, Sweatshops, and the Net," *Political Communication* 17, no. 4 (October 2000): 403–7, https://doi.org/10.1080/10584600050179022.
119. "Consumer Activism: A Growing Threat to Corporate Reputation," *Commetric* (blog), November 29, 2019, https://commetric.com/2019/11/29/consumer-activism-a-growing-threat-to-corporate-reputation/.
120. Hanlin Li et al., "How Do People Change Their Technology Use in Protest? Understanding," *Proceedings of the ACM on Human-Computer Interaction* 3, no. CSCW (November 7, 2019): 87:1–87:22, https://doi.org/10.1145/3359189.
121. "Americans Are Changing Their Relationship with Facebook," *Pew Research Center* (blog), September 5, 2018, https://www.pewresearch.org/fact-tank/2018/09/05/americans-are-changing-their-relationship-with-facebook/.
122. Paul Lewis and Erin McCormick, "How an Ex-YouTube Insider Investigated Its Secret Algorithm," *The Guardian*, February 2, 2018, sec. Technology,

https://www.theguardian.com/technology/2018/feb/02/youtube-algorithm-election-clinton-trump-guillaume-chaslot.
123. Issie Lapowsky, "YouTube Will Crack Down on Toxic Videos, But It Won't Be Easy," *Wired*, January 25, 2019, https://www.wired.com/story/youtube-recommendations-crackdown-borderline-content/.
124. Julia Alexander, "YouTube Claims Its Crackdown on Borderline Content Is Actually Working," *The Verge*, December 3, 2019, https://www.theverge.com/2019/12/3/20992018/youtube-borderline-content-recommendation-algorithm-news-authoritative-sources.
125. Thomas Germain, "GoodRx Stops Sending Prescription Data to Facebook," Consumer Reports, March 6, 2020, https://www.consumerreports.org/health-privacy/goodrx-stops-sending-prescription-data-to-facebook/.
126. Cecilia Kang, "Consumer Groups Accuse Facebook of Duping Children," *The New York Times*, February 21, 2019, sec. Technology, https://www.nytimes.com/2019/02/21/technology/facebook-children-privacy.html.
127. "Human Rights in the Age of Artificial Intelligence," Access Now, November 2018, https://www.accessnow.org/cms/assets/uploads/2018/11/AI-and-Human-Rights.pdf.
128. Artificial Intelligence, Privacy International, accessed May 11, 2020, https://privacyinternational.org/learn/artificial-intelligence.
129. "Digital Lab," Consumer Reports, accessed May 11, 2020, https://digital-lab.consumerreports.org/.
130. "Our Members," Consumers International, accessed May 11, 2020, https://www.consumersinternational.org/members/.
131. "Will Artificial Intelligence Make Us Less Free?," American Civil Liberties Union, accessed April 5, 2020, https://www.aclu.org/issues/privacy-technology/will-artificial-intelligence-make-us-less-free.
132. "Tell Facebook: Time for a Civil Rights Audit," ColorOfChange.org, accessed April 5, 2020, http://act.colorofchange.org/sign/facebook_audit/.
133. "The Toronto Declaration," Toronto Declaration, accessed April 5, 2020, https://www.torontodeclaration.org/.
134. Spandana Singh, "Why Policy Makers Need Technologists," New America, May 22, 2018, https://www.newamerica.org/millennials/dm/why-policy-makers-need-technologists/.
135. "A Future of Failure? The Flow of Technology Talent into Government and Civil Society—A Report," Freedman Consulting, LLC, 2016, https://www.fordfoundation.org/media/1893/afutureoffailure.pdf.
136. Corinne Cath, "Governing Artificial Intelligence: Ethical, Legal and Technical Opportunities and Challenges," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376, no. 2133 (November 28, 2018): 20180080, https://doi.org/10.1098/rsta.2018.0080.
137. "Office for Artificial Intelligence," GOV.UK, accessed May 11, 2020, https://www.gov.uk/government/organisations/office-for-artificial-intelligence.
138. United States Digital Service, accessed May 11, 2020, https://usds.gov/.
139. "Public Interest Technology," New America, accessed May 11, 2020, https://www.newamerica.org/public-interest-technology/.
140. TechCongress, accessed May 11, 2020, https://www.techcongress.io.
141. Spandana Singh, "Why Policy Makers Need Technologists," New America, May 22, 2018, https://www.newamerica.org/millennials/dm/why-policy-makers-need-technologists/.
142. "Mozilla Raises Concerns Over Facebook's Lack of Transparency," The Mozilla Blog, January 31, 2019, https://blog.mozilla.org/blog/2019/01/31/mozilla-raises-concerns-over-facebooks-lack-of-transparency.
143. "Annual Self-Assessment Reports of Signatories to the Code of Practice on Disinformation 2019," European Commission, October 29, 2019, https://ec.europa.eu/digital-single-market/en/news/annual-self-assessment-reports-signatories-code-practice-disinformation-2019.
144. "Open Letter: Facebook, Do Your Part Against Disinformation," The Mozilla Blog, February 11, 2019, https://blog.mozilla.org/blog/2019/02/11/open-letter-facebook-do-your-part-against-disinformation.
145. "Facebook and Google: This Is What an Effective Ad Archive API Looks Like," The Mozilla Blog, March 27, 2019, https://blog.mozilla.org/blog/2019/03/27/facebook-and-google-this-is-what-an-effective-ad-archive-api-looks-like.

146. "OECD Principles on AI," *OECD*, May 22, 2019, https://www.oecd.org/going-digital/ai/principles/
147. Masumi Koizumi, "G20 Ministers Agree on Guiding Principles for Using Artificial Intelligence," *The Japan Times Online*, June 8, 2019, https://www.japantimes.co.jp/news/2019/06/08/business/g20-ministers-kick-talks-trade-digital-economy-ibaraki-prefecture/.
148. "White Paper on Artificial Intelligence: A European approach to excellence and trust," *European Commission*, Feb 19, 2020, https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf.
149. "OECD AI's Live Repository of over 260 AI Strategies & Policies," OECD.AI, accessed April 30, 2020, https://oecd.ai/dashboards.
150. "AI in the UK: Ready, Willing and Able?," Artificial Intelligence Committee, April 16, 2017, https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/10002.htm.
151. "Translation: Chinese Expert Group Offers 'Governance Principles' for 'Responsible AI,'" New America, June 17, 2019, https://www.newamerica.org/cybersecurity-initiative/digichina/blog/translation-chinese-expert-group-offers-governance-principles-responsible-ai/.
152. "AI Ethics Principles," Department of Industry, Science, Energy and Resources (Department of Industry, Science, Energy and Resources, September 2, 2019), https://www.industry.gov.au/data-and-publications/building-australias-artificial-intelligence-capability/ai-ethics-framework/ai-ethics-principles.
153. "Artificial Intelligence," Infocomm Media Development Authority, accessed April 5, 2020, http://www.imda.gov.sg/infocomm-media-landscape/SGDigital/tech-pillars/Artificial-Intelligence.
154. Kate Fazzini, "Europe's Huge Privacy Fines against Marriott and British Airways Are a Warning for Google and Facebook," *CNBC*, July 10, 2019, https://www.cnbc.com/2019/07/10/gdpr-fines-vs-marriott-british-air-are-a-warning-for-google-facebook.html.
155. Adam Satariano, "Google Is Fined $57 Million Under Europe's Data Privacy Law," *The New York Times*, January 21, 2019, sec. Technology, https://www.nytimes.com/2019/01/21/technology/google-europe-gdpr-fine.html.
156. "GDPR Article 22: Automated Individual Decision-Making, Including Profiling," EUGDPRAcademy, accessed May 11, 2020, https://advisera.com/eugdpracademy/gdpr/automated-individual-decision-making-including-profiling/.
157. Ibid.
158. Kalliopi Spyridaki, "GDPR and AI: Friends, Foes or Something in Between?" SAS Europe, accessed April 5, 2020, https://www.sas.com/en_us/insights/articles/data-management/gdpr-and-ai--friends--foes-or-something-in-between-.html.
159. "EDPB Response to the MEP Sophie in't Veld's Letter on Unfair Algorithms," European Data Protection Board, January 31, 2020, https://edpb.europa.eu/our-work-tools/our-documents/letters/edpb-response-mep-sophie-int-velds-letter-unfair-algorithms_en.
160. "GDPR Article 5: Principles Relating to Processing of Personal Data," EUGDPRAcademy, accessed May 11, 2020, https://advisera.com/eugdpracademy/gdpr/principles-relating-to-processing-of-personal-data/.
161. Ibid.
162. "GDPR Article 35: Data Protection Impact Assessment," EUGDPRAcademy, accessed May 11, 2020, https://advisera.com/eugdpracademy/gdpr/data-protection-impact-assessment/.
163. Daisuke Wakabayashi, Katie Benner, and Steve Lohr, "Justice Department Opens Antitrust Review of Big Tech Companies," *The New York Times*, July 23, 2019, sec. Technology, https://www.nytimes.com/2019/07/23/technology/justice-department-tech-antitrust.html.
164. Mike Ananny and Kate Crawford, "Seeing without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability," *New Media*, Dec 13, 2016, https://doi.org/10.1177/1461444816676645.
165. "White Paper on Artificial Intelligence: A European approach to excellence and trust," *European Commission*, February 19, 2020,

https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf.
166. Ibid.
167. Andrew D. Selbst and Solon Barocas, "The Intuitive Appeal of Explainable Machines," *SSRN Scholarly Paper* (Rochester, NY: Social Science Research Network, March 2, 2018), https://doi.org/10.2139/ssrn.3126971.
168. Mozilla, "Facebook and Google: This Is What an Effective Ad Archive API Looks Like," The Mozilla Blog, March 27, 2019, https://blog.mozilla.org/blog/2019/03/27/facebook-and-google-this-is-what-an-effective-ad-archive-api-looks-like.
169. "EU to Invest 1.5 Billion Euros in AI to Catch up with US, Asia," *Reuters*, April 25, 2018, https://www.reuters.com/article/us-eu-artificialintelligence-idUSKBN1HW1WL.
170. "Germany Plans 3 Billion in AI Investment: Government Paper," *Reuters*, November 13, 2018, https://www.reuters.com/article/us-germany-intelligence-idUSKCN1NI1AP.
171. Charlie Taylor, "EU Presents New Plans for AI and Data as It Looks to Catch-up with US, China," *The Irish Times*, February 19, 2020, https://www.irishtimes.com/business/technology/eu-presents-new-plans-for-ai-and-data-as-it-looks-to-catch-up-with-us-china-1.4178526.
172. Thomas H. Davenport, "China Is Overtaking the U.S. as the Leader in Artificial Intelligence," *MarketWatch*, March 7, 2019, https://www.marketwatch.com/story/china-is-overtaking-the-us-as-the-leader-in-artificial-intelligence-2019-02-27.
173. Arvind Narayanan, "How to Recognize AI Snake Oil," Princeton University, accessed May 11, 2020, https://www.cs.princeton.edu/~arvindn/talks/MIT-STS-AI-snakeoil.pdf.
174. "Draft Guidelines for AI Procurement," Office for Artificial Intelligence, GOV.UK, September 20, 2019, https://www.gov.uk/government/publications/draft-guidelines-for-ai-procurement/draft-guidelines-for-ai-procurement.
175. "Data Ethics Framework," GOV.UK, August 30, 2018, https://www.gov.uk/government/publications/data-ethics-framework.
176. Dillon Reisman, Jason Schultz, Kate Crawford, and Meredith Whittaker, "Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability," AI Now Institute, April 2018, https://ainowinstitute.org/aiareport2018.pdf.
177. "Declaration of Cities Coalition for Digital Rights," Cities Coalition for Digital Rights, accessed May 11, 2020, https://citiesfordigitalrights.org/#declaration.
178. DeepSpeech, Mozilla, accessed May 11, 2020, https://github.com/mozilla/deepspeech.
179. George Roter, "Sharing Our Common Voices – Mozilla Releases the Largest to-Date Public Domain Transcribed Voice Dataset," The Mozilla Blog, February 28, 2019, https://blog.mozilla.org/blog/2019/02/28/sharing-our-common-voices-mozilla-releases-the-largest-to-date-public-domain-transcribed-voice-dataset.
180. "Can Your Holiday Gift Spy on You?," Mozilla, November 20, 2019, https://foundation.mozilla.org/en/blog/can-your-holiday-gift-spy-you/.

Appendix A

What follows below is a full size version of Mozilla's AI Theory of Change. This theory of change seeks to enable Mozilla and our allies to take both coordinated and decentralized action in a shared direction, towards collective impact on trustworthy AI.

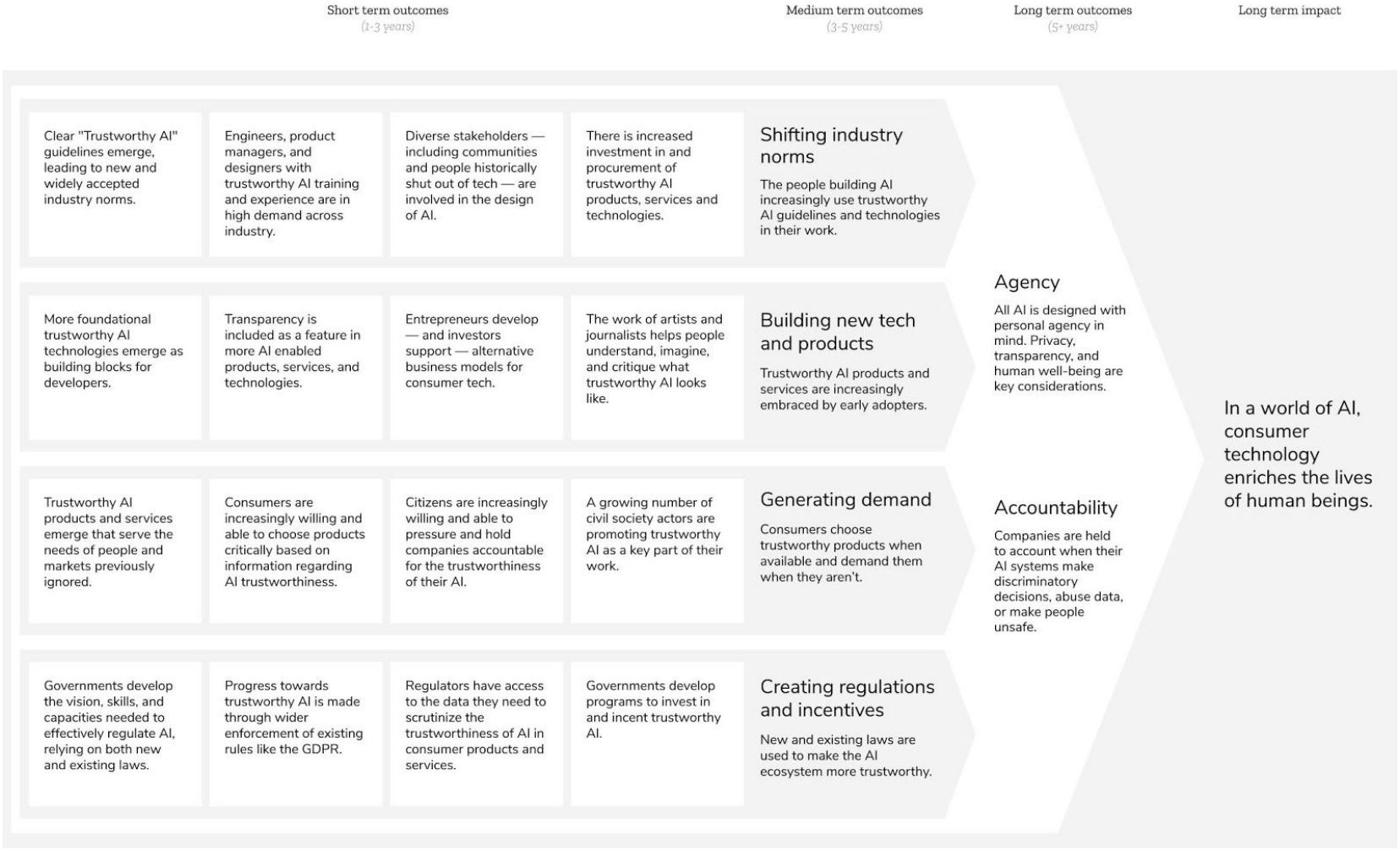
You can find it, and other supporting strategy documents regularly updated and maintained on the Mozilla AI wiki: https://wiki.mozilla.org/Foundation/AI

In particular, this theory of change seeks to define:

- Tangible changes in the world we and others will pursue (aka long term outcomes)
- Strategies that we and others might use to pursue these outcomes
- Results we will hold ourselves accountable to

Mozilla will both measure itself — and map the work of collaborators and others — against this theory of change over coming years. This will give us a way to track our progress towards trustworthy AI.

AI Theory of Change



References

Ananny, Mike, and Kate Crawford. "Seeing without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability." *New Media & Society*, December 13, 2016. https://doi.org/10.1177/1461444816676645.

Baack, Stefan and Madeleine Maxwell. "Alternative Data Governance Approaches: Global Landscape Scan and Analysis." Mozilla Foundation. September 2020. https://foundation.mozilla.org/en/initiatives/data-futures/data-for-empowerment/.

Buolamwini, Joy, and Timnit Gebru. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification." In *Conference on Fairness, Accountability and Transparency*, 77–91. PMLR, 2018. http://proceedings.mlr.press/v81/buolamwini18a.html.

Bullert, B. J. "Progressive Public Relations, Sweatshops, and the Net." *Political Communication* 17, no. 4 (October 1, 2000): 403–7. https://doi.org/10.1080/10584600050179022.

Burrell, Jenna. "How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms." *Big Data & Society* 3, no. 1 (January 5, 2016): 205395171562251. https://doi.org/10.1177/2053951715622512.

Cath, Corinne. "Governing Artificial Intelligence: Ethical, Legal and Technical Opportunities and Challenges." *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376, no. 2133 (November 28, 2018): 20180080. https://doi.org/10.1098/rsta.2018.0080.

Costanza-Chock, Sasha. *Design Justice: Community-Led Practices to Build the Worlds We Need*. Cambridge, MA: MIT Press, 2020.

Delacroix, Sylvie, and Neil Lawrence. "Bottom-Up Data Trusts: Disturbing the 'One Size Fits All' Approach to Data Governance." SSRN Scholarly Paper. Rochester, NY: Social Science Research Network, October 12, 2018. https://doi.org/10.2139/ssrn.3265315.

DiResta, Renee. "Computational Propaganda: If You Make It Trend, You Make It True." *The Yale Review* 106, no. 4 (2018): 12–29. https://doi.org/10.1111/yrev.13402.

Dheeru Dua and Casey Graff. UCI Machine Learning Repository. Irvine, California: University of California, School of Information and Computer Science. http://archive.ics.uci.edu/ml. Accessed July 20, 2020.

Elish, M. C. and Tim Hwang. “An AI Pattern Language.” New York: Data & Society, 2017. https://www.datasociety.net/pubs/ia/AI_Pattern_Language.pdf. Accessed May 11, 2020.

European Commission. “White Paper on Artificial Intelligence: A European approach to excellence and trust.” Brussels: European Commission, 2020. Accessed May 11, 2020. https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf.

Fiesler, Casey, Natalie Garrett, and Nathan Beard. “What Do We Teach When We Teach Tech Ethics? A Syllabi Analysis.” In *Proceedings of the 51st ACM Technical Symposium on Computer Science Education*, 289–295. SIGCSE ’20. Portland, OR, USA: Association for Computing Machinery, 2020. https://doi.org/10.1145/3328778.3366825.

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

Gray, Mary L. and Siddharth Suri. *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass*. Boston: Houghton Mifflin Harcourt, 2019.

Gurses, Seda, and Joris V. J. van Hoboken. 2017. “Privacy After the Agile Turn.” *SocArXiv*, May 2, 2017. https://doi.org/10.31235/osf.io/9gy73.

Hagiu, Andrei, and Julian Wright. “Multi-Sided Platforms.” SSRN Scholarly Paper. Rochester, NY: Social Science Research Network, March 19, 2015. https://doi.org/10.2139/ssrn.2794582.

Hardjono, Thomas, and Alex Pentland. “Data Cooperatives: Towards a Foundation for Decentralized Personal Data Management.” *ArXiv:1905.08819 [Cs]*, May 21, 2019. http://arxiv.org/abs/1905.08819.

Jensen, Mark A., Vincent Ferretti, Robert L. Grossman, and Louis M. Staudt. “The NCI Genomic Data Commons as an Engine for Precision Medicine.” *Blood* 130, no. 4 (July 27, 2017): 453–59. https://doi.org/10.1182/blood-2017-03-735654.

Jin, Ginger Zhe. “Artificial Intelligence and Consumer Privacy.” SSRN Scholarly Paper. Rochester, NY: Social Science Research Network, January 1, 2018. https://papers.ssrn.com/abstract=3112040.

Jobin, Anna, Marcello Ienca, and Effy Vayena. “The Global Landscape of AI Ethics Guidelines.” *Nature Machine Intelligence* 1, no. 9 (September 2019): 389–99. https://doi.org/10.1038/s42256-019-0088-2.

Li, Hanlin, Nicholas Vincent, Janice Tsai, Jofish Kaye, and Brent Hecht. “How Do People Change Their Technology Use in Protest? Understanding.” *Proceedings of the ACM on Human-Computer Interaction* 3, no. CSCW (November 7, 2019): 87:1–87:22. https://doi.org/10.1145/3359189.

Madaio, Michael, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. “Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI.” In *CHI Conference on Human Factors in Computing Systems*. ACM, 2020. https://www.microsoft.com/en-us/research/publication/co-designing-checklists-to-understand-organizational-challenges-and-opportunities-around-fairness-in-ai/.

Noble, Safiya Umoja. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York, NY: New York University Press, 2018.

Ortolano, Leonard, and Anne Shepherd. “Environmental Impact Assessment: Challenges and Opportunities.” *Impact Assessment* 13, no. 1 (March 1995): 3–30. https://doi.org/10.1080/07349165.1995.9726076.

Sandvig, Christian, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. “Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms.” Paper presented to “Data and Discrimination: Converting Critical Concerns into Productive Inquiry,” a preconference at the 64th Annual Meeting of the International Communication Association, May 22, 2014; Seattle, WA, USA. http://www-personal.umich.edu/~csandvig/research/Auditing%20Algorithms%20--%20Sandvig%20--%20ICA%202014%20Data%20and%20Discrimination%20Preconference.pdf.

Selbst, Andrew D., and Solon Barocas. “The Intuitive Appeal of Explainable Machines.” SSRN Scholarly Paper. Rochester, NY: Social Science Research Network, March 2, 2018. https://doi.org/10.2139/ssrn.3126971.

Simonyan, Karen, Andrea Vedaldi, and Andrew Zisserman. “Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps.” *ArXiv:1312.6034 [Cs]*, April 19, 2014. http://arxiv.org/abs/1312.6034.

Speicher, Till, Muhammad Ali, Giridhari Venkatadri, Filipe Ribeiro, George Arvanitakis, Fabrício Benevenuto, Krishna P Gummadi, Patrick Loiseau, and Alan Mislove. “Potential for Discrimination in Online Targeted Advertising.” In *FAT 2018 - Conference on Fairness, Accountability, and Transparency*, 81:1–15. New York, United States, 2018. https://hal.archives-ouvertes.fr/hal-01955343.

Ticona, Julia, Alexandra Mateescu, and Alex Rosenblat. “Beyond Disruption: How Tech Shapes Labor Across Domestic Work & Ridehailing.” New York: Data & Society, 2018. Accessed May 11, 2020. https://datasociety.net/library/beyond-disruption/.

Webb, Amy. *The Big Nine: How The Tech Titans and Their Thinking Machines Could Warp Humanity*. New York: PublicAffairs/Hachette Book Group, 2019.

West, Sarah Myers, Meredith Whittaker, and Kate Crawford., “Discriminating Systems: Gender, Race, and Power in AI.” New York: AI Now Institute, 2019. https://ainowinstitute.org/discriminatingsystems.pdf. Accessed May 11, 2020.

Wu, David J. “Fully Homomorphic Encryption: Cryptography’s Holy Grail.” *XRDS: Crossroads, The ACM Magazine for Students* 21, no. 3 (March 27, 2015): 24–29. https://doi.org/10.1145/2730906.

Wykstra, Stephanie. “Developing a More Diverse AI.” *Stanford Social Innovation Review*, 17, no. 1 (Winter 2019). https://ssir.org/articles/entry/developing_a_more_diverse_ai.