Communicating About the Social Implications of AI: A FrameWorks Strategic Brief

August 2021

Lindsey Conklin, PhD, Senior Researcher
Emilie L'Hôte, PhD, Director of Research
Patrick O'Shea, PhD, Senior Researcher
Michelle Smirnova, PhD, Research Fellow

In partnership with The John D. and Catherine T. MacArthur Foundation

FRAME WORKS

Communicating About the Social Implications of AI

A FrameWorks Strategic Brief

Introduction

"Don't be evil" has survived as Google's unofficial motto long after the organization officially switched to the more proactive "Do the right thing."[1] The original motto reflects widespread assumptions about technology and artificial intelligence. People think of technology as an objective product of science and numbers, free from human tendency for bias and error, where accurate data lead to accurate results. They also believe that tech is always intended for *good*, and that good intent leads to positive impact. In turn, people assume that AI could only have negative consequences if an individual designer or user *intentionally* decided to "be evil." People don't see how bias can exist in different parts of the AI process or how the use of AI in domains like policing, child welfare, or health care ends up reproducing and reinforcing existing structures of power and oppression in society. In turn, people underestimate the need as a society to manage AI.

The main goal of this brief is to further explore the deep assumptions that make it hard for the American public to understand the social implications of AI. To understand how the public perceives AI, we narrowed in on one specific technology, predictive algorithms, which have consistent problems such as amplifying inequities across different social issues. We then chose three domains in which predictive algorithms are commonly used—policing, child welfare, and health care—to see how thinking and resulting framing strategies may need to differ in the context of different issues. This brief is part of a broader project that will develop a comprehensive strategy to effectively communicate about the social implications of AI, conducted by the FrameWorks Institute in partnership with The John D. and Catherine T. MacArthur Foundation's Technology in the Public Interest (TPI) program.

Based on detailed research[2], we show that advocates and activists hoping to gain public support for social change on AI-related issues are up against five main obstacles—five common ways of thinking that make it hard for people to understand AI and its implications. These are shaped in part by sci-fi discourse about the "robots taking over," but are also grounded in more deeply ingrained beliefs and assumptions about intelligence, humanity, and technology.

First, we show that the public does not have a clear grasp of what AI is and isn't, and primarily thinks of "AI" as a stand-in for any new, impressive technology. Second, while people recognize the importance of what data is used in AI, they don't have a deeper understanding of what predictive algorithms are or how they work. Third, members of the public are unclear about the relationship between humans and AI, fixating on AI as a human replacement instead of asking whether and how AI should become incorporated into different areas of society. Fourth, while people can sometimes see that AI can make mistakes, they don't see how algorithms reproduce systemic societal biases and instead view AI-related problems as dependent on access to AI based on income level. Finally, while people seem open to some government intervention in the field of AI, they are focused on the role of "bad actors," and thus underestimate the need for collective action. Throughout this brief, we offer recommendations on how advocates and activists can address obstacles and leverage openings in public thinking to build support for needed policies.

What are We Trying to Communicate?

To develop an effective strategy for communicating about the social implications of AI, it's necessary to first identify a set of key ideas to convey. To do this, FrameWorks researchers conducted 11 interviews and a feedback session with researchers and advocates working on the social implications of AI, reviewing relevant literature on the issue. We paid particular attention to the social implications of AI across three specific domains: policing, child welfare, and health care. Below, we summarize the key ideas that emerged from this process, which represent the core points that need to be effectively communicated and the solutions for which the field wants to build support through communications.

What is artificial intelligence (AI)?

— Artificial intelligence (AI) refers to computer systems that can perform tasks that normally require human intelligence.

— Modern applications of AI often use large datasets to automate and augment human decision-making.

What are predictive algorithms and how do they work?

— Predictive algorithms analyze current or historical data to make predictions about future events.

— Predictive algorithms identify hidden trends and patterns in data.

— Predictive algorithms are shaped by human decision-making at all levels.

— Predictive algorithms are used to aid decision-making in a range of systems and settings, including policing, child welfare, and health care.

What are the current problems with how predictive algorithms are used?

— Predictive algorithms are often based on flawed or biased data, which affects their accuracy and efficacy.

— The ways that predictive algorithms work reflect the identities, values, and goals of those who design them.

— There is little accountability and oversight by regulating entities such as the government around the use of predictive algorithms.

— The use of predictive algorithms can reflect, reproduce, and exacerbate systemic biases in society.

What needs to happen ensure that predictive algorithms are used responsibly?

— Address the underlying sources of bias in systems and institutions that use predictive algorithms.

— Diversify the AI field with respect to gender, race, class, disability status, and other characteristics.

— Ensure that predictive algorithms are transparent about design and data used and open to public critique.

— Regulate (or prohibit) specific applications of AI on the basis of their potential to do harm.

Public Thinking About the Social Implications of AI

To explore the public's thinking about the social implications of AI, researchers at FrameWorks conducted 25 one-on-one, two-hour-long cognitive interviews with members of the American public. These interviews were analyzed to identify the deep, implicit ways of thinking that the public uses in contemplating the social implications of AI. We chose to focus on three different domains in which AI is commonly used: policing, child welfare, and health care. This helped us to understand the ways in which public thinking about the social implications of AI is similar or different across domains.

AI is utilized within systems of power and oppression—for example, AI augments existing discriminatory and racist policing policies. Our analysis is therefore attentive to how racism, discrimination, capitalism, and classism shape—or do not shape—people's thinking about the social implications of AI.

Based on this research, we identified five obstacles that communicators face in conveying the key ideas described above. For each obstacle, we start with what people already understand before exploring how people's presumed assumptions affect thinking on its core aspects. We offer initial recommendations on how to respond to the obstacles and leverage the openings, which communicators can start using immediately, with the important caveat that further research is needed to identify more specific, evidence-based framing strategies that can move public thinking to better understand the social implications of AI.

Obstacle #1: The public doesn't have a clear grasp of what AI is and what it isn't.

What people *don't* get

When the public thinks of examples of AI, they sometimes land on things that are AI—such as health care imaging tests like MRIs—but in general, AI is thought of as a term for "cutting-edge" or new, impressive technology.

People widely associate technology with devices, so when they think about AI, they tend to focus on innovative, "smart" devices, regardless of whether a given device actually relies on AI. The AI industry itself often plays off this misunderstanding when it comes to venture funding[3] and describes any new technology as AI—even when it does not use it—to receive larger investments. Some conspicuous examples include Fitbit watches that track movement and glucose monitors that track blood sugar, and similarly, the public often assumes that devices with "smart" technology such as Apple's Siri, Amazon's Alexa, or Google Assist centrally rely on AI.

The public is also prone to think of AI as a replacement for human beings, which means that when people see smart devices like these, they think of them as personal assistants. The devices are perceived to mimic humans by learning individual preferences and tailoring responses to meet a person's needs just like an "artificial" personal assistant. Notably, when thinking of AI, Alexa and Fitbit were more apparent to the public than sci-fi scenarios about "robots taking over" (though this is also a perception of AI).

What's not immediate for people when thinking about AI are the less-tangible algorithms and problems associated with them, as well as some of the uses that pervade daily life, such as Google's search engine or social media algorithms that filter what news items people see. This is a perfect illustration of the "AI Effect," coined by Pamela McCorduck in her 1979 book, *Machines Who Think*,[4] in which she argues that once an AI technology becomes familiar, it is no longer thought of as AI.

What does this mean for the field?

The assumptions that *all* new, innovative technologies rely on AI—and that AI is *only* used in new, innovative technological advances—make it harder in three ways for people to see what negative impact AI can have on society:

— When people assume that more forms of technology rely on AI than actually do, they are likely to focus their attention and energy on identifying the risks of technology that might be new but unrelated to AI (e.g., devices that track blood glucose levels).

— When people disregard established forms of AI, such as the Google Search algorithm or social media algorithms, it makes it harder for them to be concerned about the negative consequences these forms of AI have on their daily lives.

— The overly expansive rhetoric used by the AI industry not only attracts investors but also makes it hard for people, including government officials unfamiliar with how the technology works, to understand the specific challenges that accompany AI.

How to address Obstacle #1

Avoid playing into the belief that "AI = innovation," as it not only obscures the basics of what AI is but might also reinforce the tech industry's AI branding tactics to gain venture funding. Instead:

- **Explain** what you mean by the term "AI" (e.g., if by "AI" you mean "predictive algorithms," clarify that at the start of your communications).
- **Give** more varied examples of what AI looks like in practice, including both established and newer examples of what AI is used for.
- **Repeat** these definitions and examples often, to build more accurate public understanding of what the term refers to.

This approach will help people understand that AI not only includes cutting-edge technological advances but is often already built into the fabric of our lives (e.g., through Google's search engine).

- **Lead** with the social issue you are trying to solve (e.g., discrimination in policing, child welfare, or health care) instead of the AI technology that is causing the problem. AI can then be positioned as an amplifier of these social issues. There are three steps to this strategy:
    - **Explain** what systemic problem you are trying to solve (e.g., in policing, child welfare, or health care).
    - **Describe** the type of AI used within this system and how it contributes to the problem.
    - **Present** solutions to address this specific problem and explain how they work, to avoid fatalism.

Explaining *first* what is at stake and *next* how AI contributes to a given social problem is likely to better convey that the social implications of AI come not just from the technology itself but also its interaction with the systems in which it operates. While more research is needed to confirm and refine this strategy, starting with the social problem is a critical way to focus attention on relevant technologies and help people see the impacts that technological innovation has on social issues.

Obstacle #2: People don't understand what predictive algorithms are or how they work.

What people *do* get

When people think about uses of AI, they focus on what goes into AI (the input) and what comes out (the output). People often use the term "fed" to describe how data is put into AI and the term "spits out" to describe what comes out. They do not see that the function or algorithm plays a key role in the use of AI or understand exactly what it is and how it works, but they do see that input affects output.

One result of this input/output model is that the public is fixated on the need for "accurate" data. People explain how it is important for the data that goes "into" AI to be accurate or input by "qualified specialists," so that whatever information comes "out" of AI can also be trusted as accurate. On the one hand, this suggests that people are aware that the AI input deeply impacts the AI output. On the other, people mostly assume that "accurate" data means it is high-quality and don't see the ways that bias can be baked into the data itself.

What does this mean for the field?

The recognition that AI needs quality data to produce solid results can serve as an entry point for people to think more expansively about what AI is, how it works, and what consequences it can have on the population.

On the other hand, it is important to expand people's understanding of what "quality" data is, to help them recognize the need for considering the context in which data are collected. The current assumption that accurate data will automatically produce acceptable results obscures the fact that societal bias is often enmeshed in data in ways that negatively impact certain groups—notably black, indigenous, and other people of color (BIPOC) and low-income communities. For example, the data used in predictive policing programs may be accurate data about *arrests,* yet the use of arrest data itself builds in societal bias. Because BIPOC individuals are more likely to be arrested due to systemic racism, this data skews predictive policing algorithms to overpolice BIPOC communities.

Because people don't yet understand what algorithms are and how they work, they rely on input-output logic to reason about AI. While the input-output logic presents an opportunity to communicate about the importance of good data, it also makes it easy to overlook major sources of bias in AI—both in terms of how algorithms are designed and the role that humans play in the process.

What people *don't* get

AI is assumed to be able to *see* and therefore know what is going on in both the present and the future. When it comes to the present, people think of predictive algorithms as surveillance, which is the ability to see and know what is currently happening. When it comes to the future, people think of predictive algorithms as divination, which is the ability to see into the future and know what is going to happen based on the information gathered in the present.

People misconstrue AI prediction as the ability to see what is happening in the present through surveillance systems meant to monitor and track, primarily through cameras. When thinking about policing, people reason that AI would instantaneously identify and enable the arrest of bad actors engaging in crime. When thinking about child welfare, participants argued that AI should be used to surveil and monitor private homes for abuse; one person took this logic to the extreme, suggesting that at-risk children be implanted with AI "chips" that would enable social workers to see what is happening to them 24/7. In the domain of health care, on the other hand, people reason that AI-powered surveillance can help "quickly" catch diseases in the body for a wide variety of tests, many of which do not rely on AI to yield results. Beyond 3D imaging like MRIs and PET scans that can utilize AI, people describe AI as offering fast analysis of tests such as bloodwork, ultrasounds, genetic testing of babies in utero, heart rate and blood pressure, oxygen levels in the lungs, and muscle mass. This line of thinking likens AI to surveillance because it offers an instantaneous snapshot of these biochemical or biological markers.

The public also think of AI prediction as the ability to *see* into the future and *prevent* bad things from happening based on information gleaned from the present. This concept of seeing as knowing creates a false sense of certainty about the future. When thinking in this way, the public sees AI less like a scientist or machine and more like a fortune teller or psychic, who they believe will be able to sense a change and predict what will happen in the future so that it can be prevented in the present. When applied to the domains of policing, child welfare, and health care, participants reasoned that the use of AI would instantaneously identify and address problems before they happened or got worse (e.g., crime, abuse, or disease). As one participant explained, the speed of AI allows it to create faster test results with scans and bloodwork and "quickly determine if there's an emergency situation in the body."

Surprisingly, concerns raised about "Big Brother" or "Minority Report" societies were mostly absent from our interviews; people tend to see increased surveillance and control over future events as a positive development. Participants were not critical about the use of AI in policing, child welfare, or health care when they assumed it was being used to "see" and "prevent" bad things from happening, or if a problem is found, to "nip it in the bud." For instance, they primarily saw surveillance systems—assumed to be AI-powered—as a great way to catch more criminals and child abusers, leading people to view surveillance positively in this context. This can be contrasted with other contexts in which people's data is collected for monetization, as on social media, that lead them to be concerned about privacy issues.

What does this mean for the field?

Public thinking about prediction as "seeing" has two impacts. First, it leads to inaccurate assumptions about what AI can do and how it could be used in society across policing, child welfare, and health care. In the case of child abusers, for example, the public assumes that AI will use surveillance technology to monitor and then stop or prevent child abuse from occurring. Second, it leads people to emphasize the "instantaneous" nature of AI that can magically predict outcomes based on information gleaned from the present. Both impacts create blinders to the most important issues with AI, distracting people from how bias is baked into AI data and algorithms.

It is also problematic that the public sees increased surveillance in a positive light and is fixated on what is assumed to be perfect future predictions. This tendency can lead people to think that the use of AI in policing or child welfare will not impact anyone who is not a "bad actor," causing them to miss the outsized impact that AI has on BIPOC or low-income groups that are more likely to be surveilled. This thinking is a barrier in getting the public to understand that AI can bring harm to certain populations because it can reflect, reproduce, and exacerbate systemic biases in society.

How to address Obstacle #2

Explain what algorithms are and how the quality of "training" data influences their design. Shedding light on the mechanisms that AI relies on can help the public see not only on what goes into AI and what comes out of it, but also the multifaceted steps that are needed to *process* data.

Make explicit the different criteria for quality data. Give examples of what "quality" data look like (e.g., data that are not collected within biased systems of power).

Explain what "prediction" means in the context of AI:

- **Clarify** that AI uses data collected from prior events to forecast—much like the weather—what could happen in the future.
- **Use** terms that evoke possibility rather than certainty, (e.g., "could," "if," "whether"), to reinforce the idea that AI generates hypotheses based on prior events but cannot literally see into the future.

Obstacle #3: The public doesn't understand the relationship between humans and AI.

What people *don't* get

Instead of asking whether and how AI should become incorporated into different areas of society, members of the public are fixated on whether AI is ultimately intended to *replace* human activities in a variety of processes. In people's minds, AI is constantly expanding its capabilities and power and may end up replacing human services. This idea of gaining power through technological capability is also what underlies more extreme, sci-fi scenarios where robots end up "taking over" the world. As people assume that AI is in perpetual progress and inevitably becomes better and more efficient over time, they logically reason that robots will end up surpassing human capabilities. In some cases, people think that AI replacing human activities is almost inevitable, while in other cases, they see replacement as an outcome to be avoided due to the serious consequences it could have on people.

Whether people think of robots replacing human activities as inevitable or to be avoided depends on their interpretation of "intelligence" as information storage and retrieval or as sentience, which includes creative thinking and the ability to feel emotions. When the public thinks of intelligence as voluminous information storage, they reason that artificial intelligence is naturally superior to human intelligence. People reason that while an individual can only hold a finite amount of knowledge, computers offer access to infinite information through the Internet, which acts as a repository from endless sources. Therefore, AI perceived as cutting-edge technology is seen as the "sum" of all information available on the planet.

When people focus on the importance of information and evidence in a social domain—notably health care and policing—they reason that the more data available, the more valid and effective findings and decisions will be. When people think of health care as a science that relies on evidence and information for accuracy, they assume that AI can advantageously replace human services by garnering more accurate evidence (such as biological measurements) and therefore make more accurate diagnoses than human judgment. For similar reasons, some participants also spoke about AI being able to identify disease trends more effectively at the population level. Along the same lines, when the public thinks that the goal of policing is to get at some objective "truth" through accumulation of proof—like detective work—they reason that AI would advantageously replace human services.

On the other hand, when people think about intelligence as sentience, which includes creative thinking and the ability to feel emotions—abilities deemed uniquely human—the public is wary of AI replacing human decision-making. This is notably the case when people focus on social activities for which care is deemed essential: any domain in which children are centrally involved[5] or health care professions in which quality of care *for* patients is shaped by how much doctors care *about* them.[6] In these cases, human intelligence is prized over

artificial intelligence because it is understood in terms of the ability to *care*, something that AI is assumed to be incapable of. When thinking both about child welfare and health care, participants often conjured images of computers and robots lacking "humanity" and being fundamentally unable to replace human caregivers. As one participant put it, the possibility of robots caring for children or sick patients seemed as unrealistic as "a dress made of water."

Similarly, when people reason that some domains of social activity require skills to think creatively and adapt to an infinite variety of uniquely human circumstances, they assume that human capabilities are irreplaceable. When participants focused on the idea that every patient and every child is different, they concluded that AI would always be inadequate to care for patients or children. They reasoned that AI could only deal with a set number of situations and that only a human being could ever be attuned to the myriad differences that arise in care settings.

What does this mean for the field?

The assumption that AI is a *replacement* for human services and judgment makes it difficult for people to see that the key issues with using AI in domains like policing, child welfare, and health care lie at the intersection of AI and humans. People already have concerns about the use of AI in certain domains of life, but they are somewhat misguided and sometimes inaccurately informed by sci-fi-like scenarios. They are stuck on the less-relevant question of whether AI can be a valid substitute for human activity altogether, instead of asking whether and how AI should become incorporated into different areas of society. Because they focus on this false "humans vs. AI" dilemma, the public doesn't currently see the forest for the trees when it comes to the social consequences of AI.

This human vs. AI dilemma also makes it hard for people to see the positive outcomes of humans working *with* AI in certain circumstances, to inform their decision-making. For example, they don't necessarily see that in health care, AI could be used as one of many data points to enhance and facilitate decision-making.

When people think of intelligence as the ability to store and retrieve information, it makes them more likely to focus on *quantity* of information rather than what is done *with* the information. It reduces AI to information retrieval, with little thinking about the bias baked into the data or the role of the algorithm itself. This line of thinking also hides the fact that AI algorithms are *designed* by and reflect the identities, values, and goals of those who create them.

How to address Obstacle #3

Highlight how humans are ***always*** centrally involved in AI, from design to use to interpretation, so that the public gets a better sense that "artificial" intelligence is never meant to be "standalone" intelligence.

Provide examples of how AI is already used to *aid* in human decision-making and explain how humans can use AI as one of many tools to achieve better outcomes in certain domains, such as health care (e.g., alerts from an AI program when a patient meets certain disease risk factors that help doctors follow up with more questions or testing). This will prevent people from getting stuck in the AI vs. human dilemma.

Obstacle #4: The public doesn't see how AI upholds an inequitable and unjust system.

What people *don't* get

People are sometimes able to see that AI, like other technology, can make mistakes. They even recognize that these mistakes can be serious and have life-and-death consequences, such as mistakes in facial recognition or medical diagnosis. Yet people are overwhelmingly convinced that any negative consequences incurred from using AI in policing, child welfare, or health care stem from "glitches and inaccuracies." They focus on glitches as the source of mistakes rather than on how AI is designed or how it is intended to function. Glitches become a way for people to explain the negative consequences of AI as exceptions to the rule, rather than the rule itself. Even when they see that problems occur on occasion, they remain convinced that AI is ultimately more accurate and less biased than humans.

What does this mean for the field?

The public's default understanding of problems with AI as "glitches" makes it hard to see that the detrimental issues in AI are built into the technology itself. The public's tendency to explain away negative consequences as the exception to the rule may interfere with deeper understandings of bias in AI. Future communications need to show people that the negative consequences of AI are the result of more intentional processes, such as how the algorithm is designed and how bias is enmeshed in the data, and avoid any language centered on "mistakes."

What people *don't* get

When people think about the social implications of AI, they focus on what money can buy, which is grounded in the broader consumerist mindset that money buys quality in all areas of life. When pushed to think about the impact that AI might have on society, people's minds go to a divide between rich and poor, and how the money people have or don't have centrally shapes the aspects of AI to which they have access or exposure.

On the one hand, AI is seen as a luxury good that can only be bought and consumed for a high price. When people focus on AI as new, impressive technology that makes life easier and provides superior outcomes, they reason that only the rich in society have the money to own and benefit from these elite AI-powered services. Participants often talked about how expensive "smart" gadgets can be and argued that only people of "intense privilege" could afford anything AI-related.

In contrast, AI is also seen as a low-cost, low-quality option for human service delivery. When people focus on the notion that AI will never be able to complete human activities that require care and creative adaptation, they reason that those in society who cannot afford the quality provided by humans will inevitably be stuck with AI as a subpar replacement. Channeling the "AI vs. humans" dilemma, participants reasoned that the poorer people are, the more likely they are to be negatively affected by AI as a cost-saving measure, with human services replaced by AI services that lack a human touch.

What does this mean for the field?

When people think about AI as related to what money can buy—either as a luxury good for the rich or a low-quality option for human service delivery among low-income populations—it leads to inaccurate understandings of why and how AI impacts historically oppressed communities. By taking a superficial, consumerist perspective on AI, people are unable to recognize that AI brings harm to certain populations because it reflects, reproduces, and exacerbates existing systemic biases in society.

People's consumerist approach to AI can also lead to fatalism: Without deeper understanding of the reasons why there is a wealth gap in the U.S. in the first place, people assume that the different experiences that the rich and the poor have with AI are just the natural way of things, and nothing can be done to challenge it.

How to address Obstacle #4

Be explicit about how AI can harm low-income communities. Absent clarity about how AI can reproduce injustice, references to income are likely to be understood through a consumerist lens.

Foreground how AI reproduces systemic biases through a range of examples showing AI's impact on people's lives. Strike a balance between more dramatic examples and more mundane examples to avoid minimizing the nuances of how AI can impact people's lives. For example, predictive policing can lead to dramatic wrongful convictions, but it can also disrupt the lives and development of children and young people in overpoliced neighborhoods.

Avoid leading with the myth you are trying to debunk (e.g., people thinking that algorithms are objective, standalone decision-making tools, when they are actually involved in their development and use). Anything people hear first tends to stick in their minds, so leading with inaccurate beliefs to show why they're inaccurate is likely to reinforce these default, unproductive ways of thinking.

Obstacle #5: People think that AI regulations are necessary to curb "bad actors."

What people *do* get

The public does believe that local, state, and federal government should regulate how AI is used, with this concept focused on monitoring the influence of ill-intentioned users or "bad actors." People describe how local and federal governments should monitor AI companies to ensure they are not using AI in unethical ways. This notion became more prominent for participants when they recognized that AI could have negative consequences on people's lives, grounded in a general view of the U.S. government as protector, which recent FrameWorks projects[7] have found to be gaining traction among the public. The public increasingly sees the government as having a responsibility to protect the population from harm and respond to their needs through effective regulations.

What does this mean for the field?

While this way of thinking needs expansion to meet the field's goals, it signals that the public can understand the need for regulations to reduce the negative impact of AI on people's lives. When the negative impact is highlighted for people, it helps them see the benefit of getting the government involved through monitoring and regulations.

What people *don't* get

Members of the public blame AI's negative impact on select bad actors, rather than on the technology itself. Members of the public have fairly positive beliefs and attitudes towards technology for two reasons: They see technology as objective, and they assume the tech industry has good intentions. These assumptions strongly shape how likely people think AI is to have a negative impact, and who or what might ultimately be responsible for that.

First, people reason that technology, including AI, is objective. Members of the public perceive technology as an "objective" product of science and numbers—an indicator of certainty that does not require interpretation. It is considered more accurate and less biased than humans, whose judgment is not only prone to mistakes but also clouded by subjective feelings and emotions. When focusing on this aspect of technology, people often contrasted the certainty and objectivity of technology to human emotions that can easily "distort" reality.

Second, the public sees technology as fundamentally positive and, in turn, assumes that the people who create technology have good intentions. In people's minds, technology is a tool intended to make life easier and accomplish tasks more quickly and efficiently. When participants discussed technology as a tool, they often compared computers and phones to more rudimentary tools like spoons or hammers. They reasoned that technology in general, and AI more specifically, are primarily intended to improve people's lives—to make a variety of daily tasks and activities (e.g., socializing, communicating, doing work, finding information, entertaining oneself) more convenient for everyone in society.

Because the public sees AI as an objective tool created from the intention to improve people's lives, they reason that any negative impact resulting from the use of AI is a deviation from the norm and stems from the person using it—not the tool itself. This way of thinking influences the public's view of what needs to happen to reduce AI's negative impact: They reason that individual users and tech corporations are primarily, if not exclusively, responsible for ensuring that AI is used ethically.

This surfaces in two ways. First, people think that one way of reducing the negative impact of AI is to educate individual users, so they know how to use AI responsibly. People describe the need for trainings to be conducted by experts, namely people who designed the technology or people who have used it for an extended period. Second, when it comes to private companies, the public puts responsibility in the hands of Big Tech to regulate its own technology and ensure its integrity. As the producers of this technology, Big Tech is assumed to be responsible for ensuring that it works properly, retains integrity, and follows proper protocols.

What does this mean for the field?

These foundational assumptions about technology make it hard for people to see the inherent issues with AI itself. The belief that technology, and therefore AI, is objective and bias free makes it especially difficult to see the role of bias in predictive algorithms: the way it is baked into the data, the algorithm, and the interpretation of the output. The belief that technology, and therefore AI, comes from fundamentally good intentions also leaves the roles and responsibilities of AI designers and researchers—and their own biases—out of the picture for the public. Moreover, this assumption makes it difficult for people to see how companies' incentives can lead to systematic problems in design and development. If problems with AI arise, they are typically blamed on ill-intentioned users, not on the technology itself or the people, companies, or markets behind them.

People's focus on individual responsibility also detracts from more systemic solutions, such as the need for specific regulations to mitigate the harmful effects of AI and to create more transparency in the field at large. Further, the idea that Big Tech as AI experts and the creators of the algorithms should simply regulate their own field and train others to behave ethically is problematic because it places too much trust in the industry to regulate itself.

How to address Obstacle #5

Lead with the change you want to see in society (e.g., decreased overpolicing of Black communities), then show how AI currently contributes to the injustices, and then end with how addressing these injustices contributes to the desired solution.

Explain what regulations would look like in practice and *how* they would protect people from negative impacts. To do this:

- **Mention** how regulations would affect the design, data, *and* users of AI, so that people do not fixate on the need to regulate only users' behavior.
- **Give** examples of regulations beyond just "monitoring" AI, to expand people's understanding of what needs to happen in the field of AI.
- **Be explicit** about the need for *public* policies and how they would make a difference. It is important to note that the word "policy" has been coopted by the AI industry to talk about their own decision-making process, rather than government's involvement in AI-related issues.
- **Draw** parallels with other types of technology that have required a public approach to regulations, such as safety belt regulations in cars.

Endnotes

1. https://gizmodo.com/google-removes-nearly-all-mentions-of-dont-be-evil-from-1826153393
2. A fuller description of the data and methods behind this research is available as a supplement to this brief.
3. See: https://www.forbes.com/sites/parmyolson/2019/03/04/nearly-half-of-all-ai-startups-are-cashing-in-on-hype/?sh=d6c78edd0221 and https://www.theverge.com/2019/3/5/18251326/ai-startups-europe-fake-40-percent-mmc-report
4. McCorduck, Pamela (2004). *Machines Who Think* (2nd ed.). A. K. Peters.
5. For a more detailed discussion of the connection between children and care, see FrameWorks Institute: *Why aren't kids a policy priority? The cultural mindsets and attitudes that keep kids off the public agenda - a FrameWorks strategic brief.* (Forthcoming).
6. In health care, notions of care focus on the degree to which doctors care about their patients, as seen in previous FrameWorks' research: *Safety is more than caring: Mapping the gaps between expert, public, and health care professional understandings of patient safety* (2017). Good or caring doctors spend a lot of time with their patients and are personally devoted to them, remember their patients' names, maintain an encyclopedic knowledge of their patients' medical histories, and are immune to pressures from the pharmaceutical industry and other sources of personal gain.
7. For a discussion on thinking around government, see FrameWorks Institute: *Why aren't kids a policy priority? The cultural mindsets and attitudes that keep kids off the public agenda (forthcoming), as well as Is culture changing in this time of social upheaval? Preliminary findings from project culture change.* (Forthcoming).

About FrameWorks

The FrameWorks Institute is a nonprofit think tank that advances the mission-driven sector's capacity to frame the public discourse about social and scientific issues. The organization's signature approach, Strategic Frame Analysis®, offers empirical guidance on what to say, how to say it, and what to leave unsaid. FrameWorks designs, conducts, and publishes multi-method, multidisciplinary framing research to prepare experts and advocates to expand their constituencies, to build public will, and to further public understanding. To make sure this research drives social change, FrameWorks supports partners in reframing, through strategic consultation, campaign design, FrameChecks®, toolkits, online courses, and in-depth learning engagements known as FrameLabs. In 2015, FrameWorks was named one of nine organizations worldwide to receive the MacArthur Award for Creative and Effective Institutions.

Learn more at www.frameworksinstitute.org

Communicating About the Social Implications of AI: A FrameWorks Strategic Brief

August 2021

Please follow standard APA rules for citation, with the FrameWorks Institute as publisher.

Conklin, L., L'Hôte, E., O'Shea, P., Smirnova, M. (2021). *Communicating about the social implications of AI: A FrameWorks strategic brief.* FrameWorks Institute.

FRAME WORKS